A comparative study on low-memory iterative solvers for FFT-based homogenization of periodic media

Nachiketa Mishra[a], Jaroslav Vondřejc[b,a], Jan Zeman[a,*]

[a]*Department of Mechanics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic.*
[b]*Institute of Scientific Computing, Technische Universität Braunschweig, Hans-Sommer-Straße 65, 381 06 Braunschweig, Germany*

Abstract

In this paper, we assess the performance of four iterative algorithms for solving non-symmetric rank-deficient linear systems arising in the FFT-based homogenization of heterogeneous materials defined by digital images. Our framework is based on the Fourier-Galerkin method with exact and approximate integrations that has recently been shown to generalize the Lippmann-Schwinger setting of the original work by Moulinec and Suquet from 1994. It follows from this variational format that the ensuing system of linear equations can be solved by general-purpose iterative algorithms for symmetric positive-definite systems, such as the Richardson, the Conjugate gradient, and the Chebyshev algorithms, that are compared here to the Eyre-Milton scheme—the most efficient specialized method currently available. Our numerical experiments, carried out for two-dimensional elliptic problems, reveal that the Conjugate gradient algorithm is the most efficient option, while the Eyre-Milton method performs comparably to the Chebyshev semi-iteration. The Richardson algorithm, equivalent to the still widely used original Moulinec-Suquet solver, exhibits the slowest convergence. Besides this, we hope that our study highlights the potential of the well-established techniques of numerical linear algebra to further increase the efficiency of FFT-based homogenization methods.

Keywords: Numerical homogenization, Fourier-Galerkin method, Fast Fourier Transform, guaranteed bounds, Richardson iteration, Conjugate gradient algorithm, Chebyshev semi-iterative method, Eyre-Milton scheme

1. Introduction

Various experimental and simulation techniques, such as serial sectioning [1], computed tomography [2], statistical reconstruction [3], or digital models [4] are currently available to characterize microstructures of heterogeneous materials in a degree of realism not possible before. When combined with the tools of homogenization theories, e.g. [5, 6, 7], these advances have made it possible to establish the structure-property relations of complex engineering materials across length scales ranging from micrometers to tens of centimeters. The scale transitions rely on the solution of the *corrector problem* – a boundary value problem defined on a representative cell of the material, typically involving periodic boundary conditions. Since the input data are provided in the form of pixel- or voxel-based geometries, the need therefore arises for efficient solvers that employ images as discretization grids. Although several finite element or finite difference solvers have been developed for this purpose (e.g. [8, 9, 10]) methods based on the Fast Fourier Transform (FFT) generally offer the best computational efficiency, because of the regular grid, the simple shape of the computational domain, and the periodic boundary conditions.

---

[*]Corresponding author

Email addresses: `nachiketa.mishra@fsv.cvut.cz` (Nachiketa Mishra), `vondrejc@gmail.com` (Jaroslav Vondřejc), `zemanj@cml.fsv.cvut.cz` (Jan Zeman)

In the field of computational micromechanics of materials, the first FFT-based homogenization solver was proposed by Moulinec and Suquet in 1994 [11] and more than twenty years later, it is still widely used because of its simplicity and computational speed. The crux of the method is to reformulate the corrector problem as an integral equation of the Lippmann-Schwinger type solved by fixed-point iterations, while taking advantage of the fact that the kernel action can be efficiently handled using FFT. Later extensions of the basic algorithm were driven by the need to (i) *accelerate its convergence* for high-contrast problems [12, 13, 14, 15, 16]; (ii) to increase *accuracy of local fields* by incorporating inclusion shapes [17], modified kernels [18, 19], or local smoothing of coefficients [20, 21]; and to (iii) *prove the convergence of approximate solutions* in the framework of spectral collocation methods [22, 23, 24], the Galerkin discretization of the non-classical Hashin-Shtrikman functionals with piecewise-constant approximation spaces [25, 26], and standard Fourier-Galerkin methods [23].

Apart from providing theoretical justification to the original scheme, the Fourier-Galerkin setting has also been found convenient from the numerical point of view. For instance, it has clarified the effects of numerical quadrature [27], and led to the development of fully explicit guaranteed error bounds on homogenized properties based on a primal-dual variational approach [23, 28], which were later shown to be more restrictive than the corresponding Hashin-Shtrikman bounds [29]. The purpose of this paper is to complement these studies by examining the performance of four low-memory iterative methods for solving linear systems associated with the Fourier-Galerkin discretizations. Our comparison involves general-purpose short-recurrence solvers, namely the Richardson scheme [30], the Conjugate gradient method [31], the Chebyshev semi-iteration [32], together with the Eyre-Milton algorithm [12] – the most efficient of the accelerated schemes developed specifically for FFT-based homogenization problems, according to the recent study [33].

Related work. Previous comparative studies on FFT-based homogenization algorithms fall into two categories. The aim of the first group of works is to compare their results with finite element solvers for material-specific applications, such as particle-reinforced composites with elasto-plastic phases [34], visco-plastic models of polycrystalline materials [35, 36, 37], or transport processes and creep in concrete-like materials [38, 39]. Results of these studies consistently reveal that FFT-based methods offer at least an order-of-magnitude improvement in the computational time while predicting very similar distributions of local fields. The second group of studies is dedicated to accelerated schemes, namely to benchmarking their computational performance for high-contrast problems [40] and to revealing that they can be derived from a common recurrence relation [33].

Contributions. Although considerable effort has been spent on benchmarking FFT-based algorithms, neither of the studies above addresses conventional iterative solvers for symmetric positive-definite systems, the applicability of which follows naturally from the Fourier-Galerkin setting [41, 23]. We aim to fill this gap while utilizing the standard techniques and results of numerical linear algebra. In particular, we discuss in detail the (i) eigenvalue distribution of the system matrix, (ii) effects of numerical integration, and reduction in (iii) algebraic errors and (iv) guaranteed bounds on homogenized properties during iterations. To the best of our knowledge, this is the first study addressing such aspects for FFT-based homogenization solvers.

Limitations. Because our goal is to provide basic insight into the behavior of the different linear solvers for FFT-based homogenization, we restrict our attention to the two-dimensional scalar linear elliptic problems with isotropic phases, moderate contrasts in coefficients, and discretizations not exceeding $\approx 3,000,000$ unknowns (corresponding to a $1,999 \times 1,999$ pixel image). We also do not provide details about the overall computational time, since all simulations were performed with an experimental Python-based code FFTHomPy, available at https://github.com/vondrejc/FFTHomPy, that is not optimized for speed. However, because

our observations are based on well-established results of numerical linear algebra, they extend directly to more involved applications of FFT-based homogenization solvers reported in the literature, as evidenced by recent contributions [42, 43, 44].

Organization of the paper. The remainder of the manuscript is organized as follows. The essentials of the Fourier-Galerkin discretization of the periodic corrector problem are briefly reviewed in Section 2 following our more detailed expositions [23, 28, 27]. In Section 3, we provide details for the linear iterative solvers considered in this study. Results of the numerical experiments are gathered in Section 4, and the paper is concluded with the summary of the most important findings in Section 5.

Notation. We will denote d-dimensional vectors and matrices by boldface letters, e.g. $\boldsymbol{a} = (a_\alpha)_{\alpha=1,\dots,d} \in \mathbb{R}^d$ or $\boldsymbol{A} = (A_{\alpha\beta})_{\alpha,\beta=1,\dots,d} \in \mathbb{R}^{d\times d}$. The Euclidean inner product will be referred to as $(\bullet, \bullet)_{\mathbb{R}^d}$ and the corresponding norm as $\|\bullet\|_{\mathbb{R}^d}$. By $\mathbb{R}^{d\times d}_{\mathrm{spd}}$, we will refer to the space of symmetric positive-definite $d \times d$ matrices.

Vectors and matrices arising from discretization on regular grids will be denoted by the bold serif font in order to highlight their special structures. In particular, for a parameter $\boldsymbol{N} \in \mathbb{N}^d$ related to the discretization along each coordinate and an index set $\mathbb{Z}^d_{\boldsymbol{N}}$ enumerating the degrees of freedom, see ahead to (7) for the exact specification, we use

$$\mathsf{a}_{\boldsymbol{N}} = \left(a_\alpha^{\boldsymbol{k}}\right)^{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha=1,\dots,d} \in \mathbb{R}^{d\times\boldsymbol{N}}, \qquad \mathsf{A}_{\boldsymbol{N}} = \left(A_{\alpha\beta}^{\boldsymbol{km}}\right)^{\boldsymbol{k},\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}}_{\alpha,\beta=1,\dots,d} \in [\mathbb{R}^{d\times\boldsymbol{N}}]^2.$$

The corresponding matrix-vector and matrix-matrix multiplication are understood as

$$(\mathsf{A}_{\boldsymbol{N}}\mathsf{a}_{\boldsymbol{N}})^{\boldsymbol{k}} = \sum_{\boldsymbol{m}\in\mathbb{Z}^d_{\boldsymbol{N}}} \boldsymbol{A}^{\boldsymbol{km}}\boldsymbol{a}^{\boldsymbol{m}}, \qquad (\mathsf{A}_{\boldsymbol{N}}\mathsf{B}_{\boldsymbol{N}})^{\boldsymbol{km}} = \sum_{\boldsymbol{n}\in\mathbb{Z}^d_{\boldsymbol{N}}} \boldsymbol{A}^{\boldsymbol{kn}}\boldsymbol{B}^{\boldsymbol{nm}} \text{ for } \boldsymbol{k}, \boldsymbol{m} \in \mathbb{Z}^d_{\boldsymbol{N}},$$

and the space $\mathbb{R}^{d\times\boldsymbol{N}}$ is endowed with the following inner product and norm

$$\left(\mathsf{a}_{\boldsymbol{N}}, \mathsf{b}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}} = \frac{1}{|\boldsymbol{N}|} \sum_{\boldsymbol{k}\in\mathbb{Z}^d_{\boldsymbol{N}}} \left(\boldsymbol{a}^{\boldsymbol{k}}, \boldsymbol{b}^{\boldsymbol{k}}\right)_{\mathbb{R}^d}, \qquad \|\mathsf{a}_{\boldsymbol{N}}\|^2_{\mathbb{R}^{d\times\boldsymbol{N}}} = \left(\mathsf{a}_{\boldsymbol{N}}, \mathsf{a}_{\boldsymbol{N}}\right)_{\mathbb{R}^{d\times\boldsymbol{N}}},$$

where $|\boldsymbol{N}| = \prod_{\alpha=1}^d N_\alpha$. The same nomenclature is used for complex-valued quantities.

2. Background

The periodic corrector problem amounts to finding the matrix $\boldsymbol{A}_{\mathrm{H}} \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ defined implicitly by the variational statement, e.g. [6, Chapter 13],

$$\left(\boldsymbol{A}_{\mathrm{H}}\boldsymbol{E}, \boldsymbol{E}\right)_{\mathbb{R}^d} = \min_{\boldsymbol{e}\in\mathscr{E}} a\left(\boldsymbol{E} + \boldsymbol{e}, \boldsymbol{E} + \boldsymbol{e}\right) = a\left(\boldsymbol{E} + \boldsymbol{e}^{(\boldsymbol{E})}, \boldsymbol{E} + \boldsymbol{e}^{(\boldsymbol{E})}\right), \tag{1}$$

where $\boldsymbol{E}$ is an arbitrary vector in $\mathbb{R}^d$. The bilinear form

$$a\left(\boldsymbol{u}, \boldsymbol{v}\right) = \int_{\mathcal{Y}} \left(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})\right)_{\mathbb{R}^d} \mathrm{d}\boldsymbol{x}, \tag{2}$$

is defined on the space of the square-integrable $\mathbb{R}^d$-valued periodic functions on the unit cell $\mathcal{Y} = \prod_{\alpha=1}^d \left(-\frac{1}{2}, \frac{1}{2}\right)$, denoted as $L^2_\#(\mathcal{Y}; \mathbb{R}^d)$, and involves the matrix-valued coefficients $\boldsymbol{A} : \mathcal{Y} \to \mathbb{R}^{d\times d}_{\mathrm{spd}}$ that satisfy

$$c_A\|\boldsymbol{v}\|^2_{\mathbb{R}^d} \leq \left(\boldsymbol{A}(\boldsymbol{x})\boldsymbol{v}, \boldsymbol{v}\right)_{\mathbb{R}^d} \leq C_A\|\boldsymbol{v}\|^2_{\mathbb{R}^d} \text{ for all } \boldsymbol{v} \in \mathbb{R}^d \text{ and almost all } \boldsymbol{x} \in \mathcal{Y}, \tag{3}$$

with $0 < c_A \leq C_A < +\infty$.

The minimization problem (1) is constrained to a subspace of $L^2_\#(\mathcal{Y}; \mathbb{R}^d)$ defined by

$$\mathcal{E} = \{\boldsymbol{v} \in L^2_\#(\mathcal{Y}; \mathbb{R}^d) : \operatorname{curl} \boldsymbol{v} = \boldsymbol{0}, \int_\mathcal{Y} \boldsymbol{v}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \boldsymbol{0}\},$$

which reflects the fact that admissible vectors can be expressed as the gradient of a $\mathcal{Y}$-periodic potential. This constraint can be conveniently enforced by the orthogonal projection operator $\mathcal{G} : L^2_\#(\mathcal{Y}; \mathbb{R}^d) \to \mathcal{E}$ given by, e.g. [6, Section 12.1] or [23, Lemma 2],

$$\mathcal{G}[\boldsymbol{v}] = \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \hat{\boldsymbol{\Gamma}}(\boldsymbol{k}) \hat{\boldsymbol{v}}(\boldsymbol{k}) \varphi_{\boldsymbol{k}}, \qquad \hat{\boldsymbol{\Gamma}}(\boldsymbol{k}) = \begin{cases} \boldsymbol{0} & \text{for } \boldsymbol{k} = \boldsymbol{0}, \\ \dfrac{\boldsymbol{k}\boldsymbol{k}^{\mathsf{T}}}{(\boldsymbol{k}, \boldsymbol{k})_{\mathbb{R}^d}} & \text{for } \boldsymbol{k} \in \mathbb{Z}^d \backslash \{\boldsymbol{0}\}, \end{cases} \tag{4}$$

where $\hat{\boldsymbol{\Gamma}}(\boldsymbol{k}) \in \mathbb{R}^{d \times d}$ are projection matrices in the Fourier space and $\hat{\boldsymbol{v}}(\boldsymbol{k}) \in \mathbb{C}^d$ is the $\boldsymbol{k}$-th Fourier coefficient of $\boldsymbol{v}$,

$$\hat{\boldsymbol{v}}(\boldsymbol{k}) = \int_\mathcal{Y} \boldsymbol{v}(\boldsymbol{x}) \varphi_{-\boldsymbol{k}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \in \mathbb{C}^d \quad \text{for } \varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \exp\left(2\pi\mathrm{i}(\boldsymbol{k}, \boldsymbol{x})_{\mathbb{R}^d}\right) \text{ with } \boldsymbol{x} \in \mathcal{Y}. \tag{5}$$

It follows from the Lax-Milgram lemma that (1) has the unique minimizer $\boldsymbol{e}^{(\boldsymbol{E})}$ for any $\boldsymbol{E} \in \mathbb{R}^d$ that satisfies the optimality conditions

$$a\left(\boldsymbol{e}^{(\boldsymbol{E})}, \boldsymbol{v}\right) = -a\left(\boldsymbol{E}, \boldsymbol{v}\right) \text{ for all } \boldsymbol{v} \in \mathcal{E}.$$

In the following sections, we will explain how to obtain computable approximation to $\boldsymbol{e}^{(\boldsymbol{E})}$ using Fourier-Galerkin methods. This procedure includes specification of the approximating functions, Section 2.1; Galerkin discretizations with approximate and exact integrations yielding the guaranteed upper bounds on the homogenized matrix $\boldsymbol{A}_\mathrm{H}$, Section 2.2; and the specification of linear systems resulting from the discretization procedure, Section 2.3.

2.1. Trigonometric polynomials

Given the order of the polynomial approximation

$$\boldsymbol{N} \in \mathbb{N}^d \text{ such that } N_\alpha \text{ is odd for all } \alpha,$$

the space of $\mathbb{R}^d$-*valued trigonometric polynomials* admits two equivalent definitions, e.g. [45, Chapter 8],

$$\mathscr{T}^d_{\boldsymbol{N}} = \left\{ \sum_{\boldsymbol{k} \in \mathbb{Z}^d_{\boldsymbol{N}}} \hat{\boldsymbol{v}}^{\boldsymbol{k}} \varphi_{\boldsymbol{k}} : \hat{\boldsymbol{v}}^{\boldsymbol{k}} = \overline{\hat{\boldsymbol{v}}^{-\boldsymbol{k}}} \in \mathbb{C}^d \right\} = \left\{ \sum_{\boldsymbol{m} \in \mathbb{Z}^d_{\boldsymbol{N}}} \boldsymbol{v}^{\boldsymbol{m}} \varphi_{\boldsymbol{N}, \boldsymbol{m}} : \boldsymbol{v}^{\boldsymbol{m}} \in \mathbb{R}^d \right\}. \tag{6}$$

These definitions involve the set of truncated frequencies

$$\mathbb{Z}^d_{\boldsymbol{N}} = \left\{ \boldsymbol{k} \in \mathbb{Z}^d : |k_\alpha| < \frac{N_\alpha}{2} \right\}, \tag{7}$$

the Fourier basis functions $\varphi_{\boldsymbol{k}}$ from (5), and the fundamental trigonometric polynomials

$$\varphi_{\boldsymbol{N}, \boldsymbol{m}}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{N}|} \sum_{\boldsymbol{k} \in \mathbb{Z}^d_{\boldsymbol{N}}} \omega_{\boldsymbol{N}}^{-\boldsymbol{m}\boldsymbol{k}} \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathcal{Y},$$

where the complex-valued coefficients

$$\omega_{\boldsymbol{N}}^{\boldsymbol{m}\boldsymbol{k}} = \exp\left(2\pi\mathrm{i} \sum_\alpha \frac{m_\alpha k_\alpha}{N_\alpha}\right) \text{ for } \boldsymbol{m}, \boldsymbol{k} \in \mathbb{Z}^d_{\boldsymbol{N}}$$

define the discrete Fourier transform (DFT).

Notice that the components $\widehat{v}^{k}$ and v^{m} in (6) associated with a trigonometric polynomial $u_N \in \mathscr{T}_N^d$ are not independent since they satisfy

$$\widehat{v}^{k} = \widehat{u}_N(k), \quad v^{m} = u_N(x_N^{m}), \quad \text{for } x_N^{m} = \left(\frac{m_\alpha}{N_\alpha} \right)_{\alpha=1,\ldots,d} \text{ and } k, m \in \mathbb{Z}_N^d,$$

where $\{x_N^{m}\}^{m \in \mathbb{Z}_N^d}$ denotes the regular grid in the real space $\mathbb{R}^d$ associated with the polynomial order N. Thus, the matrices $\widehat{\mathbf{v}}_N = (\widehat{v}^{k})^{k \in \mathbb{Z}_N^d}$ and $\mathbf{v}_N = (v^{m})^{m \in \mathbb{Z}_N^d}$ can be mapped on each other with the help of DFT

$$\widehat{\mathbf{v}}_N = \mathbf{F}_N \mathbf{v}_N \in \mathbb{C}^{d \times N}, \qquad\qquad \mathbf{v}_N = \mathbf{F}_N^{-1} \widehat{\mathbf{v}}_N \in \mathbb{R}^{d \times N},$$

where the matrices implementing the forward and inverse DFT are given by

$$\mathbf{F}_N = \frac{1}{|N|} \left(\delta_{\alpha\beta} \omega_N^{-mk} \right)_{\alpha,\beta=1,\ldots,d}^{m,k \in \mathbb{Z}_N^d} \in \left[\mathbb{C}^{d \times N} \right]^2, \qquad \mathbf{F}_N^{-1} = \left(\delta_{\alpha\beta} \omega_N^{mk} \right)_{\alpha,\beta=1,\ldots,d}^{m,k \in \mathbb{Z}_N^d} \in \left[\mathbb{C}^{d \times N} \right]^2; \qquad (8)$$

$\delta_{\alpha\beta}$ stands for the Kronecker delta equal to one for $\alpha = \beta$, and to zero otherwise.

In the following sections, we shall make repeated use of the *discretization operator* $\mathcal{I}_N : C_\#(\mathcal{Y}; \mathbb{R}^d) \to \mathbb{R}^{d \times N}$,

$$\mathcal{I}_N[v] = \left(v(x_N^{k}) \right)^{k \in \mathbb{Z}_N^d},$$

that assigns the values of a continuous periodic function v at the regular grid $\{x_N^{k}\}^{k \in \mathbb{Z}_N^d}$ to the vector from $\mathbb{R}^{d \times N}$. The operator $\mathcal{I}_N$ establishes a scalar product-preserving one-to-one map between $\mathscr{T}_N^d$ and $\mathbb{R}^{d \times N}$, with the inverse $\mathcal{I}_N^{-1} : \mathbb{R}^{d \times N} \to \mathscr{T}_N^d$, and can thus be used with advantage to evaluate the action of the bilinear form (2) on trigonometric polynomials.

2.2. Galerkin approximations and bounds

The *conforming* finite-dimensional space on which the Galerkin approximations will be performed consists of curl-free trigonometric polynomials with zero mean

$$\mathscr{E}_N = \mathcal{G}[\mathscr{T}_N^d] = \mathscr{E} \cap \mathscr{T}_N^d.$$

The homogenized matrix $A_{\mathrm{H},N} \in \mathbb{R}_{\mathrm{spd}}^{d \times d}$, associated with the *Galerkin approximation* (Ga) to the corrector problem (1), then satisfies

$$\left(A_{\mathrm{H},N} E, E \right)_{\mathbb{R}^d} = \min_{e_N \in \mathscr{E}_N} a\left(E + e_N, E + e_N \right) = a\left(E + e_N^{(E)}, E + e_N^{(E)} \right) \qquad (9)$$

for arbitrary $E \in \mathbb{R}^d$. The most straightforward approach to the *exact integration* in (9) utilizes the Plancherel theorem in the Fourier domain and leads to dense matrix representations, e.g. [46] or [28, Section 6]. However, sparsity is recovered when the integration is transferred to a double grid, as shown in [28, Section 6]:

$$a\left(u_N, v_N \right) = \left(\mathbf{A}_{2N-1} \mathbf{u}_{N,2N-1}, \mathbf{v}_{N,2N-1} \right)_{\mathbb{R}^{d \times (2N-1)}}, \qquad (10)$$

with $\mathbf{u}_{N,2N-1} = \mathcal{I}_{2N-1}[u_N]$, $\mathbf{v}_{N,2N-1} = \mathcal{I}_{2N-1}[v_N]$, and the block-diagonal matrix $\mathbf{A}_{2N-1}$ provided by

$$\mathbf{A}_{2N-1}^{km} = \delta^{km} \sum_{n \in \mathbb{Z}_{2N-1}^d} \omega_{2N-1}^{kn} \widehat{A}(n), \qquad (11)$$

where $\widehat{A}(n) \in \mathbb{C}^{d \times d}$ denotes the n-th Fourier coefficient of A, recall (5). For pixel- or voxel-wise constant coefficients, $\mathbf{A}_{2N-1}$ can be assembled efficiently by FFT, see [27, Section 4] for details.

Because the exact integration is relatively involved, we also introduce a simpler strategy based on the *Galerkin approximation with numerical integration* (GaNi). It employs the trapezoidal rule leading to discretization-dependent bilinear forms $\tilde{a}_{\boldsymbol{N}} : \mathscr{T}_{\boldsymbol{N}}^d \times \mathscr{T}_{\boldsymbol{N}}^d \to \mathbb{R}$,

$$a\big(\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}\big) \approx \tilde{a}_{\boldsymbol{N}}\big(\boldsymbol{u}_{\boldsymbol{N}}, \boldsymbol{v}_{\boldsymbol{N}}\big) = \frac{1}{|\boldsymbol{N}|} \sum_{\boldsymbol{k} \in \mathbb{Z}_{\boldsymbol{N}}^d} \boldsymbol{A}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}}) \boldsymbol{u}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}}) \boldsymbol{v}_{\boldsymbol{N}}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}}) = \big(\widetilde{\mathbf{A}}_{\boldsymbol{N}} \mathbf{u}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d \times \boldsymbol{N}}},$$

where $\mathbf{u}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{u}_{\boldsymbol{N}}]$, $\mathbf{v}_{\boldsymbol{N}} = \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{v}_{\boldsymbol{N}}]$, and the block-diagonal matrix $\widetilde{\mathbf{A}}_{\boldsymbol{N}} = \big(\delta^{\boldsymbol{km}} \boldsymbol{A}(\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}})\big)^{\boldsymbol{k}, \boldsymbol{m} \in \mathbb{Z}_{\boldsymbol{N}}^d}$ collects the coefficients at the grid points. In analogy to (1) and (9), the corresponding corrector problem amounts to finding the matrix $\widetilde{\boldsymbol{A}}_{\mathrm{H}, \boldsymbol{N}} \in \mathbb{R}_{\mathrm{spd}}^{d \times d}$ defined by

$$\big(\widetilde{\boldsymbol{A}}_{\mathrm{H}, \boldsymbol{N}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = \min_{\boldsymbol{e}_{\boldsymbol{N}} \in \mathscr{E}_{\boldsymbol{N}}} \tilde{a}_{\boldsymbol{N}}\big(\boldsymbol{E} + \boldsymbol{e}_{\boldsymbol{N}}, \boldsymbol{E} + \boldsymbol{e}_{\boldsymbol{N}}\big) = \tilde{a}_{\boldsymbol{N}}\big(\boldsymbol{E} + \widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}, \boldsymbol{E} + \widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}\big). \tag{12}$$

Furthermore, combining the two conforming Galerkin schemes (9) and (12) with the variational principle (1), one immediately obtains the *guaranteed upper bounds* on the homogenized matrix

$$\big(\boldsymbol{A}_{\mathrm{H}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} \leq \big(\boldsymbol{A}_{\mathrm{H}, \boldsymbol{N}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} \leq \big(\widetilde{\boldsymbol{A}}_{\mathrm{H}, \boldsymbol{N}}^{\mathrm{bound}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d} = a\big(\boldsymbol{E} + \widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}, \boldsymbol{E} + \widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}\big) \quad \forall \boldsymbol{E} \in \mathbb{R}^d, \tag{13}$$

where the matrix $\widetilde{\boldsymbol{A}}_{\mathrm{H}, \boldsymbol{N}}^{\mathrm{bound}} \in \mathbb{R}_{\mathrm{spd}}^{d \times d}$ follows from the action of the bilinear form a from (2) to the GaNi minimizers $\widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}$ evaluated with the exact integration formula (10), cf. [28, Section 6] and [27, Section 4]. Note that analogous arguments establish guaranteed *lower* bounds on $\boldsymbol{A}_{\mathrm{H}}$ according to the dual variational principle, finally leading to fully explicit *discretization* error bounds on the approximate solutions. In the present paper, however, we shall work with the upper bounds (13) only; an interested reader is referred to [28, 27] for full details.

2.3. Linear systems

The fully discrete versions of the optimality conditions for Ga (9) and GaNi (12) follow from suitable applications of the discretization operators, resulting in

$$\big(\mathbf{A}_{2\boldsymbol{N}-1} \mathbf{e}_{\boldsymbol{N}, 2\boldsymbol{N}-1}^{(\boldsymbol{E})}, \mathbf{v}_{\boldsymbol{N}, 2\boldsymbol{N}-1}\big)_{\mathbb{R}^{d \times (2\boldsymbol{N}-1)}} = -\big(\mathbf{A}_{\boldsymbol{N}, 2\boldsymbol{N}-1} \mathbf{E}_{2\boldsymbol{N}-1}, \mathbf{v}_{2\boldsymbol{N}-1}\big)_{\mathbb{R}^{d \times (2\boldsymbol{N}-1)}}, \tag{14a}$$

$$\big(\widetilde{\mathbf{A}}_{\boldsymbol{N}} \widetilde{\mathbf{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d \times \boldsymbol{N}}} = -\big(\widetilde{\mathbf{A}}_{\boldsymbol{N}} \mathbf{E}_{\boldsymbol{N}}, \mathbf{v}_{\boldsymbol{N}}\big)_{\mathbb{R}^{d \times \boldsymbol{N}}}, \tag{14b}$$

for all $\mathbf{v}_{\boldsymbol{N}, 2\boldsymbol{N}-1} \in \mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N}-1}$ and $\mathbf{v}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$, with

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N}-1} &= \mathcal{I}_{2\boldsymbol{N}-1}[\mathscr{E}_{\boldsymbol{N}}] \subset \mathbb{R}^{d \times (2\boldsymbol{N}-1)}, & \mathbb{E}_{\boldsymbol{N}} &= \mathcal{I}_{\boldsymbol{N}}[\mathscr{E}_{\boldsymbol{N}}] \subset \mathbb{R}^{d \times \boldsymbol{N}}, \\
\mathbf{e}_{\boldsymbol{N}, 2\boldsymbol{N}-1}^{(\boldsymbol{E})} &= \mathcal{I}_{2\boldsymbol{N}-1}[\boldsymbol{e}_{\boldsymbol{N}}^{(\boldsymbol{E})}] \in \mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N}-1}, & \widetilde{\mathbf{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})} &= \mathcal{I}_{\boldsymbol{N}}[\widetilde{\boldsymbol{e}}_{\boldsymbol{N}}^{(\boldsymbol{E})}] \in \mathbb{E}_{\boldsymbol{N}}, \\
\mathbf{E}_{2\boldsymbol{N}-1} &= \mathcal{I}_{2\boldsymbol{N}-1}[\boldsymbol{E}] \in \mathbb{R}^{d \times (2\boldsymbol{N}-1)}, & \mathbf{E}_{\boldsymbol{N}} &= \mathcal{I}_{\boldsymbol{N}}[\boldsymbol{E}] \in \mathbb{R}^{d \times \boldsymbol{N}},
\end{aligned}$$

so, for example, $\mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N}-1}$ contains the nodal values of trigonometric polynomials from $\mathscr{E}_{\boldsymbol{N}}$ at the double grid points $\{\boldsymbol{x}_{\boldsymbol{N}}^{\boldsymbol{k}}\}^{\boldsymbol{k} \in \mathbb{Z}_{2\boldsymbol{N}-1}^d}$.

To obtain the systems of linear equations defined by the optimality conditions (14), we need to enforce the constraints $\mathbf{v}_{\boldsymbol{N}, 2\boldsymbol{N}-1} \in \mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N}-1}$ and $\mathbf{v}_{\boldsymbol{N}} \in \mathbb{E}_{\boldsymbol{N}}$ by means of suitable projections. Thanks to the properties of trigonometric polynomials, Section 2.1, such *discrete orthogonal projections* follow directly from the continuous version (4):

$$\mathbf{G}_{\boldsymbol{N}, \boldsymbol{M}} = \mathbf{F}_{\boldsymbol{M}}^{-1} \widehat{\mathbf{G}}_{\boldsymbol{N}, \boldsymbol{M}} \mathbf{F}_{\boldsymbol{M}} \quad \text{with} \quad \big(\widehat{\mathbf{G}}_{\boldsymbol{N}, \boldsymbol{M}}\big)^{\boldsymbol{km}} = \begin{cases} \mathbf{0} & \text{for } \boldsymbol{k}, \boldsymbol{m} \in \mathbb{Z}_{\boldsymbol{M}}^d \setminus \mathbb{Z}_{\boldsymbol{N}}^d \\ \delta^{\boldsymbol{km}} \widehat{\boldsymbol{\Gamma}}(\boldsymbol{k}), & \text{for } \boldsymbol{k}, \boldsymbol{m} \in \mathbb{Z}_{\boldsymbol{N}}^d, \end{cases} \tag{15}$$

so that, e.g. $\mathbf{G}_{N,2N-1} : \mathbb{R}^{d\times(2N-1)} \to \mathbb{E}_{N,2N-1}$; we abbreviate $\mathbf{G}_{N,N}$ to $\mathbf{G}_N$ in what follows. Now, the linear systems corresponding to Ga and GaNi arise as, cf. [23, Proposition 12] and [27, Corollary 28],

$$\mathbf{G}_{N,2N-1}\mathbf{A}_{2N-1}\mathbf{e}_{N,2N-1}^{(E)} = -\mathbf{G}_{N,2N-1}\mathbf{A}_{2N-1}\mathbf{E}_{2N-1}, \tag{16a}$$

$$\mathbf{G}_N\widetilde{\mathbf{A}}_N\widetilde{\mathbf{e}}_N^{(E)} = -\mathbf{G}_N\widetilde{\mathbf{A}}_N\mathbf{E}_N. \tag{16b}$$

Once the solutions to these linear systems have been obtained, the guaranteed upper bounds from (13) can be made explicit:

$$\big(\boldsymbol{A}_{\mathrm{H},N}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \big(\mathbf{A}_{2N-1}(\mathbf{E}_{2N-1}+\mathbf{e}_{N,2N-1}^{(E)}),\mathbf{E}_{2N-1}+\mathbf{e}_{N,2N-1}^{(E)}\big)_{\mathbb{R}^{d\times(2N-1)}}, \tag{17a}$$

$$\big(\widetilde{\boldsymbol{A}}_{\mathrm{H},N}^{\mathrm{bound}}\boldsymbol{E},\boldsymbol{E}\big)_{\mathbb{R}^d} = \big(\mathbf{A}_{2N-1}\mathcal{R}_{N,2N-1}[\mathbf{E}_N+\widetilde{\mathbf{e}}_N^{(E)}],\mathcal{R}_{N,2N-1}[\mathbf{E}_N+\widetilde{\mathbf{e}}_N^{(E)}]\big)_{\mathbb{R}^{d\times(2N-1)}}, \tag{17b}$$

where the prolongation operator $\mathcal{R}_{N,2N-1} = \mathcal{I}_{2N-1} \circ \mathcal{I}_N^{-1}$ maps $\mathbb{R}^{d\times N}$ to $\mathbb{R}^{d\times(2N-1)}$ through the intermediate space of trigonometric polynomials $\mathscr{T}_N^d$.

3. Solvers

In order to avoid a profusion of notation, we shall refer to the linear systems (16) in a unified way as

$$\mathbf{C}\mathbf{x} = \mathbf{b} \text{ for } \mathbf{x} \in \mathbb{E}, \tag{18}$$

so that, e.g., for the GaNi variant, $\mathbf{C} = \mathbf{G}_N\widetilde{\mathbf{A}}_N$, $\mathbf{x} = \widetilde{\mathbf{e}}_N^{(E)}$, $\mathbf{b} = -\mathbf{G}_N\widetilde{\mathbf{A}}_N\mathbf{E}_N$, and $\mathbb{E} = \mathbb{E}_N$. We shall also abbreviate $\big(\bullet,\bullet\big)_{\mathbb{R}^{d\times N}}$ or $\big(\bullet,\bullet\big)_{\mathbb{R}^{d\times(2N-1)}}$ to $\big(\bullet,\bullet\big)_2$, $\|\bullet\|_{\mathbb{R}^{d\times N}}$ or $\|\bullet\|_{\mathbb{R}^{d\times(2N-1)}}$ to $\|\bullet\|_2$, and $\mathbf{E}_N$ or $\mathbf{E}_{2N-1}$ to $\mathbf{E}$ when there is no risk of confusion.

Notice that in general, matrix $\mathbf{C}$ in (18) is non-symmetric and highly rank-deficient. However, as first demonstrated by Vondřejc et al. [41], $\mathbf{C}$ acts as a symmetric positive-definite matrix for vectors from the subspace $\mathbb{E}$ and satisfies

$$c_A\|\mathbf{x}\|_2^2 \le \big(\mathbf{C}\mathbf{x},\mathbf{x}\big)_2 \le C_A\|\mathbf{x}\|_2^2 \text{ for } \mathbf{x} \in \mathbb{E},$$

where c_A and C_A are the bounds on the material coefficients $\boldsymbol{A}$ from (3); the condition number of $\mathbf{C}$ on $\mathbb{E}$ can be estimated from above *independently of discretization* by $\kappa = C_A/c_A$ (we invite an interested reader to refer to Section 4.1, where these properties are demonstrated on concrete examples).

Because multiplication by $\mathbf{C}$ can be performed efficiently using the FFT in $\mathcal{O}(|\boldsymbol{N}|\log|\boldsymbol{N}|)$ operations, recall the definition of projection operator in (15) and (8), the problem (18) turns out to be well-suited to conventional iterative algorithms for symmetric positive definite systems, provided that all iterates generated by the algorithm, $\mathbf{x}_m$, remain in $\mathbb{E}$.

The remainder of this section is devoted to four such iterative solvers, covering general-purpose short-recurrence iterative algorithms, Sections 3.1–3.3; and a special-purpose solver, Section 3.4. In order to keep our exposition compact but self-contained, for each algorithm we present its pseudo-code; discuss the complexity of a single iteration, error estimates, and conformity of the iterates $\mathbf{x}_m \in \mathbb{E}$; and briefly comment on their applications to Ga- and GaNi-based formulations in earlier studies.

3.1. Richardson iteration

The Richardson iteration [30] belongs to the group of stationary iterative methods. It searches for a fixed point of the mapping

$$\mathbf{x}_m = \mathbf{x}_{m-1} - \omega(\mathbf{C}\mathbf{x}_{m-1} - \mathbf{b}) \text{ for } m \in \mathbb{N} \text{ and } \mathbf{x}_0 \in \mathbb{E} \tag{19}$$

by means of Algorithm 1.

Algorithm 1 (Richardson iteration). *Input:* $\mathbf{C}$, $\mathbf{b}$, $\mathbf{x}_0 = \mathbf{0}$, c_A, C_A, *and* ϵ

$\omega \longleftarrow \frac{2}{c_A + C_A}$ *(optimal choice)*

For $m = 1, 2, \cdots$

$\qquad \mathbf{x}_m \longleftarrow \mathbf{x}_{m-1} - \omega\left(\mathbf{C}\mathbf{x}_{m-1} - \mathbf{b}\right)$

until $\|\mathbf{x}_m - \mathbf{x}_{m-1}\|_2 \leq \epsilon\|\mathbf{E}\|_2$

return $\mathbf{x}_m$

A current implementation, one iteration of the Richardson solver involves three vectors, $\mathbf{x}_m$, $\mathbf{x}_{m-1}$, and $\mathbf{b}$, and one FFT-based matrix-vector multiplication $\mathbf{C}\mathbf{x}_{m-1}$ that dominates its computational cost. The convergence analysis of the Richardson scheme is available, e.g., in [47] or [48, Section 4.2], where it is shown that the optimal choice of the iteration parameter

$$\omega = \frac{2}{c_A + C_A} \tag{20}$$

provides the error bound for the m-th iteration in the form

$$\left\|\mathbf{x}_m - \mathbf{x}\right\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^m \left\|\mathbf{x}_0 - \mathbf{x}\right\|_2 \text{ for } m \in \mathbb{N}_0. \tag{21}$$

To see that all iterates are conforming to $\mathbb{E}$, we rewrite the iterations from Algorithm 1 as

$$\mathbf{x}_m = \mathbf{x}_{m-1} - \omega\mathbf{r}_{m-1} \text{ for } m \in \mathbb{N} \text{ and } \mathbf{x}_0 \in \mathbb{E},$$

where $\mathbf{r}_{m-1} = \mathbf{C}\mathbf{x}_{m-1} - \mathbf{b}$ is the $(m-1)$-th residual vector which is located in $\mathbb{E}$, because both the matrix $\mathbf{C}$ and vector $\mathbf{b}$ involve the discrete projection operator; recall the linear systems (16). Hence, the approximate solutions are conforming, $\mathbf{x}_m \in \mathbb{E}$, provided that the initial guess satisfies $\mathbf{x}_0 \in \mathbb{E}$.

Rather interestingly, the Richardson iteration applied to the GaNi system (16b) yields the original variant of the Moulinec-Suquet scheme [11]; the optimal choice of the iteration parameter (20) then corresponds to the results of convergence studies reported in, e.g. [12, 14, 15].[1] Applicability for Ga-based discretization has been recently reported by Monchiet [29].

3.2. Conjugate gradient method

The conjugate gradient method [31] constructs the iterates by projecting the system (16) on to a sequence of Krylov subspaces generated by the initial residual vector,

$$\mathbb{K}_m = \text{span}\left\{\mathbf{r}_0, \mathbf{C}\mathbf{r}_0, \mathbf{C}^2\mathbf{r}_0, \cdots, \mathbf{C}^{m-1}\mathbf{r}_0\right\} \text{ for } m \leq \dim(\mathbb{E}), \tag{22}$$

utilizing a coupled two-term recurrence defined by Algorithm 2.

Algorithm 2 (Conjugate gradients). *Input:* $\mathbf{C}$, $\mathbf{b}$, $\mathbf{x}_0$, *and* ϵ

$\qquad \mathbf{r}_0 \longleftarrow \mathbf{b} - \mathbf{C}\mathbf{x}_0$

$\qquad \mathbf{p}_0 \longleftarrow \mathbf{r}_0$

$\qquad$ **For** $m = 1, 2, \cdots$

$\qquad\qquad \alpha_{m-1} \longleftarrow \left(\mathbf{r}_{m-1}, \mathbf{r}_{m-1}\right)_2 / \left(\mathbf{C}\mathbf{p}_{m-1}, \mathbf{p}_{m-1}\right)_2$

$\qquad\qquad \mathbf{x}_m \longleftarrow \mathbf{x}_{m-1} + \alpha_{m-1}\mathbf{p}_{m-1}$

$\qquad\qquad \mathbf{r}_m \longleftarrow \mathbf{r}_{m-1} - \alpha_{m-1}\mathbf{C}\mathbf{p}_{m-1}$

$\qquad\qquad \beta_{m-1} \longleftarrow \left(\mathbf{r}_m, \mathbf{r}_m\right)_2 / \left(\mathbf{r}_{m-1}, \mathbf{r}_{m-1}\right)_2$

$\qquad\qquad \mathbf{p}_m = \mathbf{r}_m + \beta_{m-1}\mathbf{p}_{m-1}$

$\qquad$ **until** $\|\mathbf{r}_m\|_2 \leq \epsilon\|\mathbf{b}\|_2$

$\qquad$ **return** $\mathbf{x}_m$

---

[1]Indeed, in the current notation, iterations of the Moulinec-Suquet algorithm are defined by the recurrence $\mathbf{x}_m = \mathbf{x}_{m-1} - \boldsymbol{\Gamma}_0\mathbf{A}(\mathbf{E} + \mathbf{x}_{m-1})$, where $\boldsymbol{\Gamma}_0 = \mathbf{G}/c_0$ stands for the (matrix) Green operator of the so-called reference problem with coefficient $c_0\mathbf{I}$. For the optimal choice $c_0 = (c_A + C_A)/2$, the two algorithms coincide.

In terms of storage, the Conjugate gradient method must keep track of three vectors, namely the solution $\mathbf{x}_m$, the residual $\mathbf{r}_m$, and the search direction $\mathbf{p}_m$. In addition, the product $\mathbf{C}\mathbf{p}_{m-1}$ must be calculated with the help of the FFT and stored to an auxiliary vector in order to keep the number of matrix-vector multiplications limited to one. Finally, notice that two scalar products are needed per a single iteration.

As shown in, e.g. [49, 50], the error in the m-th iteration is bounded by

$$\left\| \mathbf{x} - \mathbf{x}_m \right\|_{\mathbf{C}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m \left\| \mathbf{x} - \mathbf{x}_0 \right\|_{\mathbf{C}}. \tag{23}$$

Since each Krylov subspace from (22) satisfies $\mathbb{K}_m \subset \mathbb{E}$, the iterates $\mathbf{x}_m$ are conforming, as first observed in [41].

To the best of our knowledge, the first heuristic applications of the conjugate gradient method to systems associated with GaNi discretization was reported independently by Brisard and Dormieux [51] and Zeman et al. [22], and justified later by Brisard and Dormieux [25] and Vondřejc et al. [41, 23]. Performance of the solver for the Ga system has been recently studied by Vondřejc [27], but no comparison with other iterative solvers has been made to date.

3.3. Chebyshev semi-iteration

The Chebyshev semi-iteration [32, 52] builds upon a generalization of the Richardson iterative formula (19)

$$\mathbf{x}_m = \mathbf{x}_{m-1} + \omega_m \mathbf{r}_{m-1} \text{ for } m \in \mathbb{N} \text{ and } \mathbf{x}_0 \in \mathbb{E}, \tag{24}$$

where ω_m is related to the roots of the Chebyshev polynomials of the first kind, shifted to the interval $[c_A, C_A]$ and suitably normalized. In our numerical experiments, ω_m is determined indirectly from a composite two-term recurrence according to Algorithm 3, because this relation proved to be the most numerically stable from the variants available in [53].

Algorithm 3 (Chebyshev iteration). *Input:* $\mathbf{C}$, $\mathbf{b}$, $\mathbf{x}_0$, c_A, C_A, and ϵ

$\quad \mathbf{r}_0 \longleftarrow \mathbf{b} - \mathbf{C}\mathbf{x}_0$
$\quad c \longleftarrow \frac{1}{2}(C_A - c_A)$
$\quad d \longleftarrow \frac{1}{2}(c_A + C_A)$
$\quad \alpha \longleftarrow \frac{1}{d}$
$\quad \beta \longleftarrow -\frac{1}{2}\left(\frac{c}{d}\right)^2$
$\quad \mathbf{r}_0 \longleftarrow \mathbf{b} - \mathbf{C}\mathbf{x}_0$
$\quad \mathbf{p}_0 \longleftarrow \mathbf{r}_0$
$\quad$ **For** $m = 1, 2, \cdots$
$\quad\quad$ **if** $m > 1$
$\quad\quad\quad \alpha \longleftarrow \left(d + \frac{\beta}{\alpha}\right)^{-1}$
$\quad\quad\quad \beta \longleftarrow -\left(\frac{c\alpha}{2}\right)^2$
$\quad\quad \mathbf{r}_m \longleftarrow \mathbf{r}_{m-1} - \alpha \mathbf{C}\mathbf{p}_{m-1}$
$\quad\quad \mathbf{x}_m \longleftarrow \mathbf{x}_{m-1} + \alpha \mathbf{p}_{m-1}$
$\quad\quad \mathbf{p}_m \longleftarrow \mathbf{r}_m - \beta \mathbf{p}_{m-1}$
$\quad$ **until** $\|\mathbf{r}_m\|_2 \leq \epsilon \|\mathbf{b}\|_2$
$\quad$ **return** $\mathbf{x}_m$

Similarly as in the case of the Conjugate gradient algorithm, the Chebyshev semi-iteration updates the triplet $(\mathbf{x}_m, \mathbf{r}_m, \mathbf{p}_m)$ at every iteration and requires only one matrix-vector product.

As discussed, e.g., in [54, Section 2], the Conjugate gradient estimate (23) holds also for the Chebyshev method and actually better reflects its true convergence behavior. By (24), all approximations $\mathbf{x}_m$ are confined by $\mathbb{E}$. Note that, to the best of our knowledge, this work represents the first application of the Chebyshev method to FFT-based homogenization.

3.4. Eyre-Milton accelerated scheme

The basic idea of the Eyre-Milton algorithm [12] is to recast the system (16) into an equivalent form with better conditioning and to solve the modified system by the Richardson scheme for the unknown in the form $\mathbf{x} = \mathbf{E}_N + \widetilde{\mathbf{e}}_N^{(E)}$. The resulting scheme appears in Algorithm 4.

Algorithm 4 (Eyre-Milton accelerated scheme). *Input:* $\mathbf{A}$, $\widehat{\mathbf{G}}$, $\mathbf{x}_0 = \mathbf{E}$, c_A, C_A, *and* ϵ

$\quad \omega \longleftarrow \sqrt{c_A C_A} \quad$ *(Optimal choice)*

$\quad \mathbf{P} \longleftarrow (\mathbf{A} + \omega\mathbf{I})^{-1} \left(\mathbf{F}^{-1}[\mathbf{I} - 2\widehat{\mathbf{G}}]\mathbf{F} \right) (\mathbf{A} - \omega\mathbf{I})$

$\quad \mathbf{q} = -2\omega(\mathbf{A} + \omega\mathbf{I})^{-1}\mathbf{x}_0$

$\quad \mathbf{For}\ m = 1,\ 2,\ \cdots$

$\quad\quad \mathbf{x}_m = \mathbf{P}\mathbf{x}_{m-1} + \mathbf{q}$

$\quad \mathbf{until}\ \|\mathbf{x}_m - \mathbf{x}_{m-1}\|_2 \leq \epsilon\|\mathbf{E}\|_2$

$\quad \mathbf{return}\ \mathbf{x}_m$

The storage requirements for the Eyre-Milton scheme are similar to those for the Richardson iteration and involve $\mathbf{x}_{m-1}$, $\mathbf{x}_m$, and $\mathbf{q}$. Multiplication with matrix $\mathbf{P}$ is more demanding than with $\mathbf{C}$, cf. (16), but its cost is dominated again by the forward and inverse FFT, and thus has the same complexity of $\mathcal{O}(|N|\log|N|)$.

The relative error bound for $\mathbf{x}_m$ [15]

$$\left\|\mathbf{x}_m - \mathbf{x}\right\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^m \left\|\mathbf{x}_m - \mathbf{x}_0\right\|_2, \tag{25}$$

is similar to the Conjugate gradient and the Chebyshev estimates (23). In addition, the iterates generated by this algorithm are not conforming in the sense that $(\mathbf{x} - \mathbf{E}) \notin \mathbb{E}$, see [33].

According to a seminal study by Moulinec and da Silva [33], the method outperforms other accelerated schemes available in the literature, namely the augmented Lagrangian formulation by Michel et al. [13, 14] and the polarization-based method by Monchiet and Bonnet [16] in the GaNi setting. However, the effect of numerical integration is very pronounced for this algorithm; quite surprisingly, we found that the method *does not converge* when matrices in Algorithm 4 correspond to Ga discretization. We attribute this behavior to the fact that coefficient matrix for GaNi, $\widetilde{\mathbf{A}}_N$, still satisfies the estimates for the continuous problem (3),

$$c_A\|\mathbf{v}_N\|^2_{\mathbb{R}^{d\times N}} \leq \left(\widetilde{\mathbf{A}}_N\mathbf{v}_N, \mathbf{v}_N\right)_{\mathbb{R}^{d\times N}} \leq C_A\|\mathbf{v}_N\|^2_{\mathbb{R}^{d\times N}} \text{ for all } \mathbf{v}_N \in \mathbb{R}^{d\times N}.$$

However, a similar condition no longer holds for $\mathbf{A}_{2N-1}$, because the exact integration impacts its eigenvalue distribution; see Section 4.1 for an explicit example.

4. Examples

Two types of unit cells appear in the comparative study. These include a single square inclusion of 36% volume fraction, Figure 1(a), and a two-dimensional cross-section of an alkali-activated ash foam sample analyzed in [55], Figure 1(b). Unless specified otherwise, the coefficients of the cell problem (3) are taken as

$$\boldsymbol{A} = \begin{cases} 1 & \text{in matrix,} \\ 100 & \text{in inclusion,} \end{cases} \qquad \boldsymbol{A} = \begin{cases} 2.6 & \text{in matrix (black),} \\ 0.026 & \text{in inclusion (white),} \end{cases} \tag{26}$$

for the square inclusion and the fly ash foam cell, respectively. The default grid of the square inclusion is set to 85×85, since the same convergence behavior was observed for finer discretizations, while the alkali-activated foam sample corresponds to a $1,999 \times 1,999$ pixel bitmap. The macroscopic field is set to $\boldsymbol{E} = [1; 0]$ and the tolerance of all four algorithms is $\epsilon = 10^{-6}$.
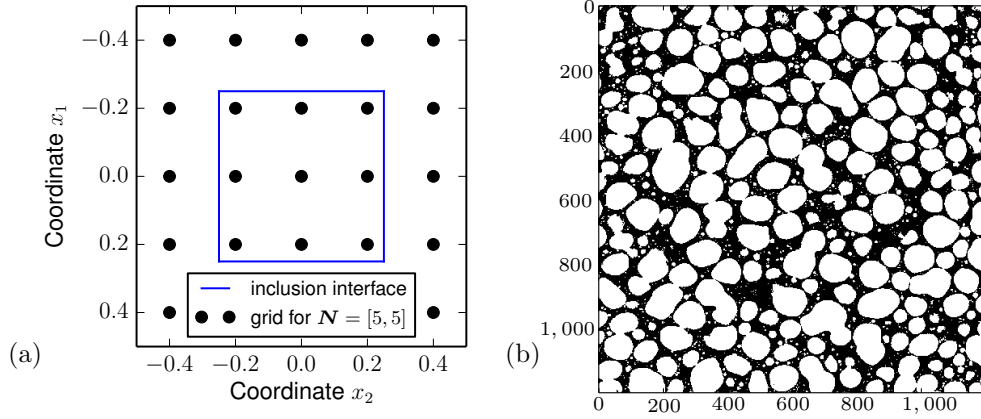
Figure 1: (a) An example of square inclusion discretized with 5×5 points and (b) $1,999 \times 1,999$ pixel bitmap of an alkali-activated fly-ash sample (courtesy of Petr Hlaváček, CTU in Prague).

4.1. Eigenvalue distribution

In order to obtain the complete eigenvalue distribution, the unit cells from Figure 1 were discretized with trigonometric polynomials of order 15×15, leading to matrix sizes $d|\boldsymbol{N}| \times d|\boldsymbol{N}| = (2 \cdot 15 \cdot 15) \times (2 \cdot 15 \cdot 15) = 450 \times 450$ for GaNi discretization and $d|2\boldsymbol{N} - \boldsymbol{1}| \times d|2\boldsymbol{N} - \boldsymbol{1}| = (2 \cdot 29 \cdot 29) \times (2 \cdot 29 \cdot 29) = 1,682 \times 1,682$ for Ga.

The cumulative distribution of eigenvalues for the two unit cell types and the two discretizations appear in Figure 2. For GaNi discretization, the matrix has a rank of 224, whereas the remaining 226 eigenvalues are zero. It further follows from the discussion in [27, Remark 30] that out of 226 zero eigenvalues, 2 correspond to the constant fields and 224 represent trigonometric polynomials of zero divergence and zero mean. For Ga, the null-space is increased by $1,232$ eigenvectors resulting from the double-grid projection (15); notice that the corresponding eigenvalues are not included in Figure 2 for better clarity.

Figure 2 illustrates how cell geometry and numerical integration influence the matrix spectra. In particular, for the square inclusion and GaNi, Figure 2(a), the eigenvalues in spectrum $[1; 100]$ form clusters. For instance no eigenvalues are present between 8 and 25 because the cumulative distribution is constant on this interval. Exact integration renders the spectrum less clustered, Figure 2(b), but the effect of geometrical irregularity is even stronger, Figure 2(c), so that the interval $[0.026; 2.6]$ is sampled uniformly with the eigenvalues of the matrix resulting from full integration applied to the fly ash foam cell, Figure 2(d). Finally, observe that irrespective of the discretization used, the non-zero eigenvalues are bounded by the coefficients of the phases according to (26); the matrix $\mathbf{C}$ is thus indeed symmetric positive-definite on $\mathbb{E}$. The extreme eigenvalues have the highest multiplicity as indicated by the corresponding jumps in the eigenvalue distribution.

Figure 3 demonstrates the effect of numerical integration on matrices of coefficients $\widetilde{\mathbf{A}}_{\boldsymbol{N}}$ (of rank 450) and $\mathbf{A}_{2\boldsymbol{N}-\boldsymbol{1}}$ (of rank 1,682) for GaNi and Ga schemes and the square inclusion. In this case, the effects of integration are even more pronounced. For GaNi scheme, the spectrum consists of two values $\{1, 100\}$, because matrix $\widetilde{\mathbf{A}}_{\boldsymbol{N}}$ contains only coefficients of the continuous problems (3) sampled on a regular grid. For Ga scheme, the exact integration (11) significantly changes the spectrum of $\mathbf{A}_{2\boldsymbol{N}-\boldsymbol{1}}$; the eigenvalues are now located within interval $[-3.89, 108.89]$. We conjecture that divergence of the Eyre-Milton scheme for Ga, reported in Section 3.4, is a direct consequence of this fact.

11
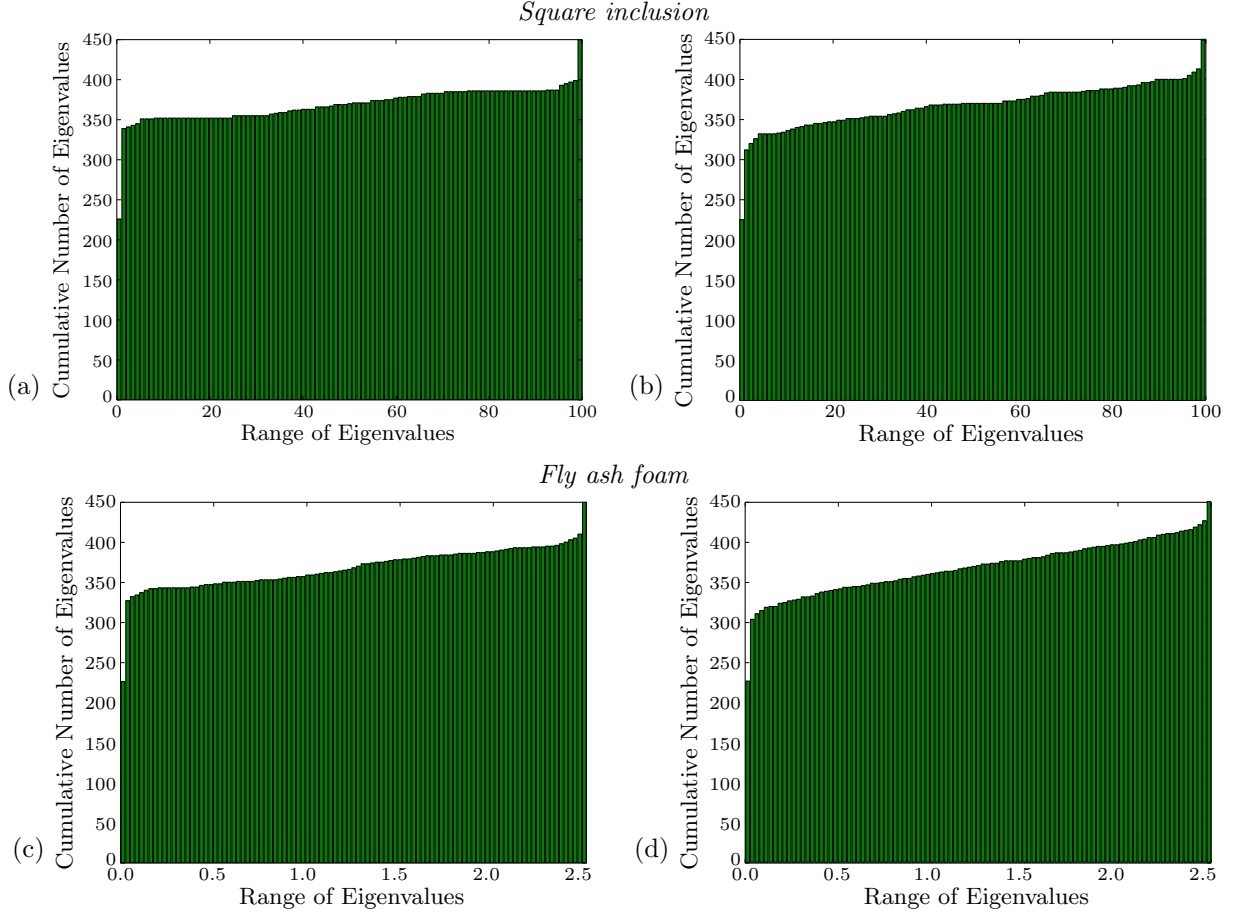
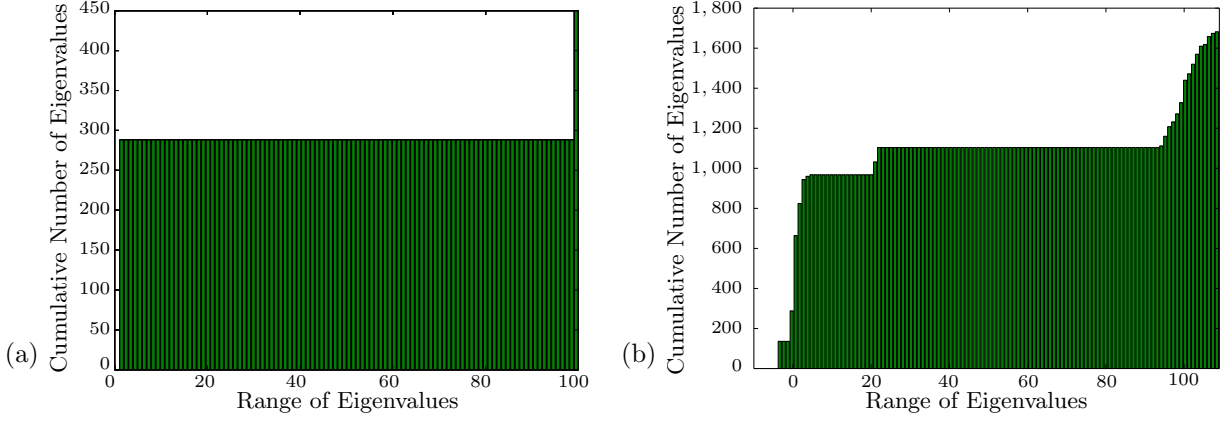Figure 2: Cumulative eigenvalue distribution for system matrices resulting from GaNi (left) and Ga (right) discretizations; non-zero eigenvalues are binned into 100 intervals.

4.2. Residual norm

We start our study on the convergence properties of FFT-based solvers by investigating the evolution of the residual norm

$$\left\|\mathbf{r}_m\right\|_2 = \left\|\mathbf{b} - \mathbf{C}\mathbf{x}_m\right\|_2 \text{ for } m \in \mathbb{N}_0$$

during iterations, see Figure 4. Results reveal that the Richardson scheme and Conjugate gradients display behavior qualitatively different from the Chebyshev method and the Eyre-Milton method. For the first two algorithms, convergence proceeds in two stages: the first stage (≈ 10 iterations) is associated with a rapid decrease in the residual norm, and then slows down in the second stage. These two stages are especially pronounced for the GaNi-discretized square inclusion, Figure 4(a), but they are clearly visible in all remaining examples.

The Chebyshev and the Eyre-Milton schemes display almost the same behavior along the whole iteration process. We attribute this behavior to the fact that the first group of algorithms initially resolves the components of the residuum vectors associated with the largest eigenvalues and then proceeds through the rest of the spectrum down to the smallest eigenvalue, cf. [56]. The second group of algorithms, on the other hand, simultaneously reduces the residuum components associated with the full spectrum, as confirmed by our computational observations that will form the basis of a separate publication.

Except for the Richardson scheme, all methods display oscillations in the convergence plots for regular geometry and GaNi, Figure 4(a), that are significantly dampened by the exact integration and/or irregular distribution of phases. These phenomena again closely follow the

Figure 3: Cumulative eigenvalue distribution for coefficient matrices resulting from GaNi (left) and Ga (right) discretization of square inclusion; eigenvalues are binned into 100 intervals.

eigenvalue distributions displayed in Figure 2. We also observe that the behavior of the Chebyshev method and Eyre-Milton scheme are almost identical for the simple geometry; for the fly ash foam microstructure, the Eyre-Milton method is more efficient and the residual norms generated by the Chebyshev method *increase* in the first iteration. At later iterations, the rate of convergence of the Richardson scheme is inferior to the remaining algorithms, which exhibit super-linear convergence in accordance with error estimates (21), (23), and (25) reported earlier in Section 3. Nevertheless, the conjugate gradient method always displays the best performance.

4.3. Guaranteed upper bound

The numerical experiments in the previous section concentrated on the residual norm because it is used as the stopping criterion in conventional solvers. Here, we shall consider the evolution of the guaranteed upper bound at the m-th iteration,

$$U_{\boldsymbol{E}}(\mathbf{x}_m) = \big(\mathbf{A}_{2\boldsymbol{N-1}}(\mathbf{E}_{2\boldsymbol{N-1}} + \mathbf{x}_m), \mathbf{E}_{2\boldsymbol{N-1}} + \mathbf{x}_m\big)_{\mathbb{R}^{d \times (2\boldsymbol{N-1})}} \text{ for } m \in \mathbb{N}_0, \qquad (27)$$

where the mapping $U_{\boldsymbol{E}} : \mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N-1}} \to \mathbb{R}$ provides the average energy in the unit cell with the distribution of local fields $(\boldsymbol{E} + \mathcal{I}_{2\boldsymbol{N-1}}^{-1}[\mathbf{x}_m])$. If the iterates $\mathbf{x}_m \in \mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N-1}}$ correspond to the Ga scheme, $U_{\boldsymbol{E}}(\mathbf{x}_m) \longrightarrow \big(\boldsymbol{A}_{\mathrm{H}, \boldsymbol{N}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d}$, recall (17a). For the GaNi scheme, the iterates $\mathbf{x}_m$ must be projected to $\mathbb{E}_{\boldsymbol{N}, 2\boldsymbol{N-1}}$ to obtain $U_{\boldsymbol{E}}(\mathcal{R}_{\boldsymbol{N}, 2\boldsymbol{N-1}}[\mathbf{G}_{\boldsymbol{N}} \mathbf{x}_m]) \longrightarrow \big(\widetilde{\boldsymbol{A}}_{\mathrm{H}, \boldsymbol{N}}^{\mathrm{bound}} \boldsymbol{E}, \boldsymbol{E}\big)_{\mathbb{R}^d}$, recall (17b). Notice that for $\mathbf{x}_0 = \mathbf{0}$, Eq. (27) corresponds to the Voigt estimate.

The behavior of individual solvers, Figure 5, agrees well with the observations made from the residual plots, Figure 4; in particular the estimates on the guaranteed upper bound generated by the Richardson scheme and by the Conjugate gradient method converge faster than the Chebyshev and Eyre-Milton schemes in the first ≈ 10 iteration, after which the bounds appear to stabilize.

In order to investigate the quality of the upper bound in more detail, we introduce the *error* in the guaranteed upper bound at the m-th iteration

$$|U_{\boldsymbol{E}}(\mathbf{x}_m) - U_{\boldsymbol{E}}(\mathbf{x}^*)| \text{ for } m \in \mathbb{N}_0$$

where $\mathbf{x}^*$ denotes the approximation to the solution of (18) obtained with the Conjugate gradient method and tolerance $\epsilon = 10^{-12}$. The results are collected in Figure 6 and demonstrate that to achieve a target accuracy of 10^{-4}, for instance, the Conjugate gradient algorithm needs less than 25 iterations for GaNi discretization and about 15 iterations for Ga, whereas the Chebyshev and Eyre-Milton methods require about 10 additional iterations to reach the same accuracy (except for fly ash foam microstructure and GaNi discretization, where the methods perform similarly). The iterates generated by the Richardson scheme deliver an accuracy of
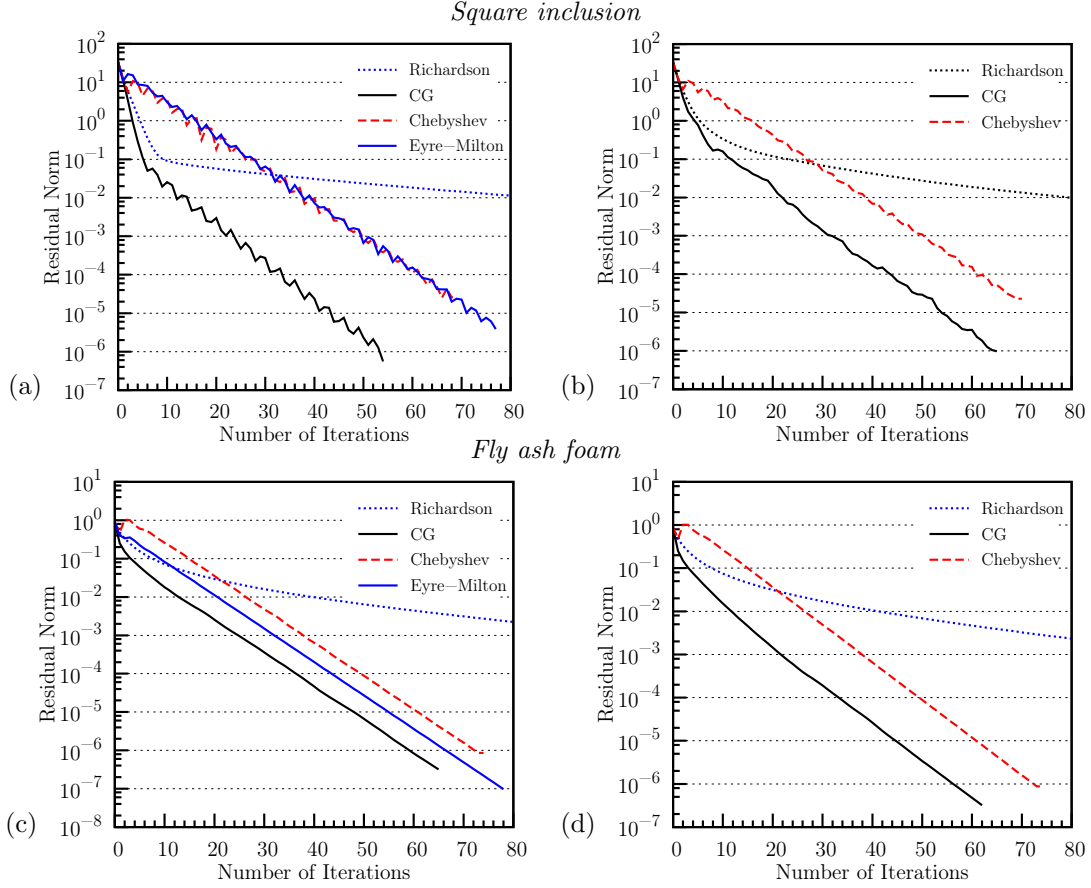
13

Figure 4: Convergence history of residual norms for GaNi (left) and Ga (right) discretizations.

10^{-4} only for Ga discretization and 70 iterations, Figure 6(b), or 60 iterations, Figure 6(d). The superior performance of the Conjugate gradient method in the Ga setting is not surprising; the upper bound (27) is exactly the energy norm that Conjugate gradients minimize over the Krylov subspaces (22), as has been recently pointed out by Vondřejc [27].

Being inspired by recent results in [27], we also find it instructive to demonstrate the effect of exact numerical integration on the tightness of the guaranteed upper bounds. Results in Figure 7 confirm that these effects are indeed significant. For square inclusion and GaNi, the bound equals to 2.793 and decreases by 20% to 2.241 by the exact integration; for the fly ash foam cell, GaNi provides a value of 0.513 that is improved to 0.469 by Ga. We believe that these results promote Ga discretization over GaNi, despite the increased computational costs due to involvement of a double grid.

4.4. Non-conformity

We shall quantify the non-conformity of iterates for GaNi-based systems by the quantities

$$\left\|\mathbf{G}_N \mathbf{x}_m + \mathbf{E}_N - \mathbf{x}_m\right\|_{\mathbb{R}^{d \times N}}, \qquad \left\|\mathbf{G}_N \mathbf{x}_m - \mathbf{x}_m\right\|_{\mathbb{R}^{d \times N}} \text{ for } m \in \mathbb{N}_0,$$

defined for the Eyre-Milton scheme and the remaining algorithms, respectively, with an obvious generalization to the Ga setting. The reason for this distinction is best seen in Figure 8(a), which highlights the difference between the non-conforming Eyre-Milton scheme and conforming solvers – the Richardson, the Conjugate gradient, and the Chebyshev algorithms. For the Eyre-Milton scheme, the non-conformity error reaches its maximum in the first iteration and progressively decreases in a non-monotone way. For conforming algorithms, the error increases with an increasing number of iterations due to the accumulation of round-off errors. Notice that
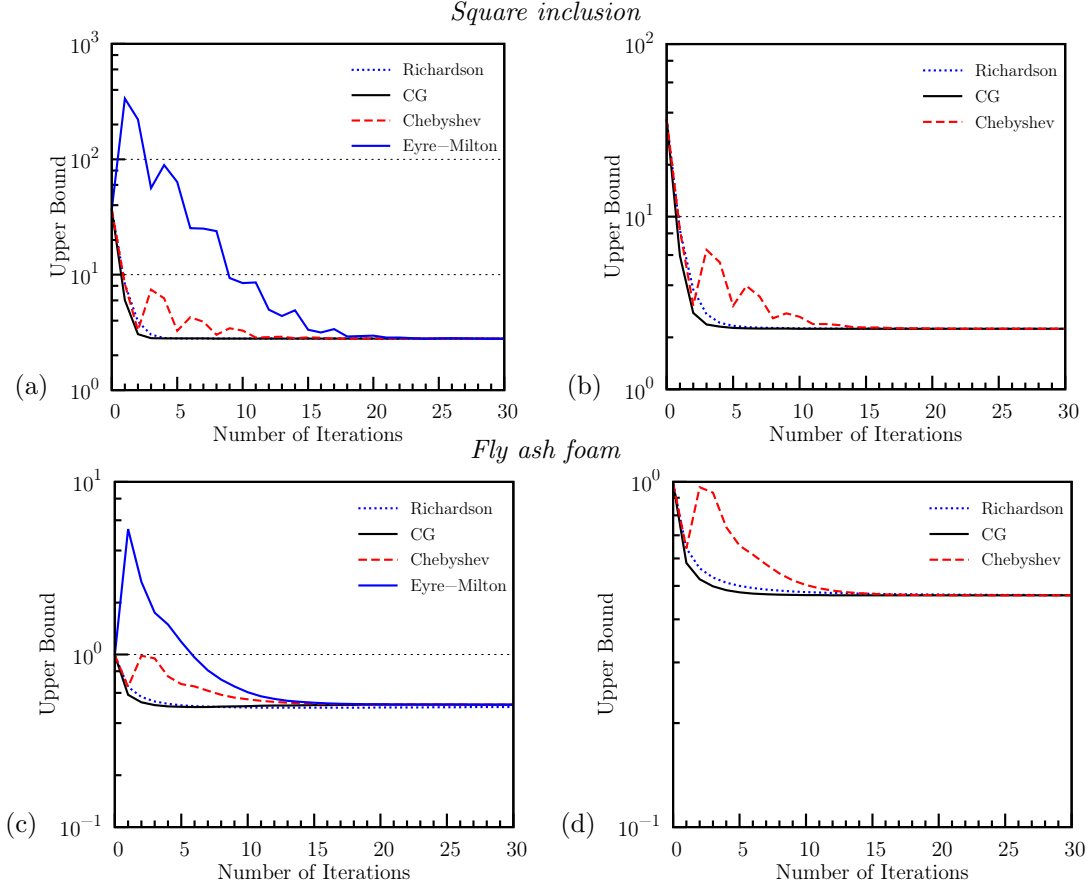
Figure 5: Convergence history of guaranteed upper bounds for GaNi (left) and Ga (right) discretizations.

although the round-off effects per iteration are smallest for the Richardson scheme, the total values are the same at convergence, Figure 8(b).

As follows from the discrete Helmholtz decomposition of $\mathbb{R}^{d \times \boldsymbol{N}}$ [28, Lemma 21], the non-conformity error of the Eyre-Milton scheme can be further split into two orthogonal components,

$$\left\|\mathsf{G}_{\boldsymbol{N}}\mathsf{x}_m + \mathsf{E}_{\boldsymbol{N}} - \mathsf{x}_m\right\|_{\mathbb{R}^{d \times \boldsymbol{N}}} = \left\|\mathsf{E}_{\boldsymbol{N}} - \langle\mathsf{x}_m\rangle\right\|_{\mathbb{R}^{d \times \boldsymbol{N}}} + \left\|\mathsf{G}_{\boldsymbol{N}}\mathsf{x}_m - \mathsf{x}_m + \langle\mathsf{x}_m\rangle\right\|_{\mathbb{R}^{d \times \boldsymbol{N}}},$$

which quantify the difference between the *mean value* of the m-iterate, defined with $\langle\mathsf{x}_m^{\boldsymbol{n}}\rangle = (\sum_{\boldsymbol{k}\in\mathbb{Z}_{\boldsymbol{N}}^d} \mathsf{x}_m^{\boldsymbol{k}})/|\boldsymbol{N}|$ for $\boldsymbol{n} \in \mathbb{Z}_{\boldsymbol{N}}^d$, and the prescribed vector $\mathsf{E}_{\boldsymbol{N}}$, and the distance from *zero-mean curl-free* vectors $\mathbb{E}_{\boldsymbol{N}}$. The relative distribution of both components, Figure 9, demonstrates that the latter component dominates the error in mean.

4.5. Relative error bound

In support of the theoretical results gathered in Section 3, our purpose is to illustrate the behavior of the relative error

$$\frac{\left\|\mathsf{x}_m - \mathsf{x}^*\right\|_{\mathsf{C}}}{\left\|\mathsf{x}_0 - \mathsf{x}^*\right\|_{\mathsf{C}}} \text{ for } m \in \mathbb{N}_0.$$

The results of this study appear in Figure 10 for the GaNi-based system and confirm the linear convergence of the Richardson scheme implied by error estimate (21) and the super-linear convergence of the remaining three algorithms, in agreement with relations (23) and (25). Notice that no results for Ga discretizations have been shown, since they are very similar to the error plots for the guaranteed upper bound, Figure 6(b,d), and lead to the same conclusions.
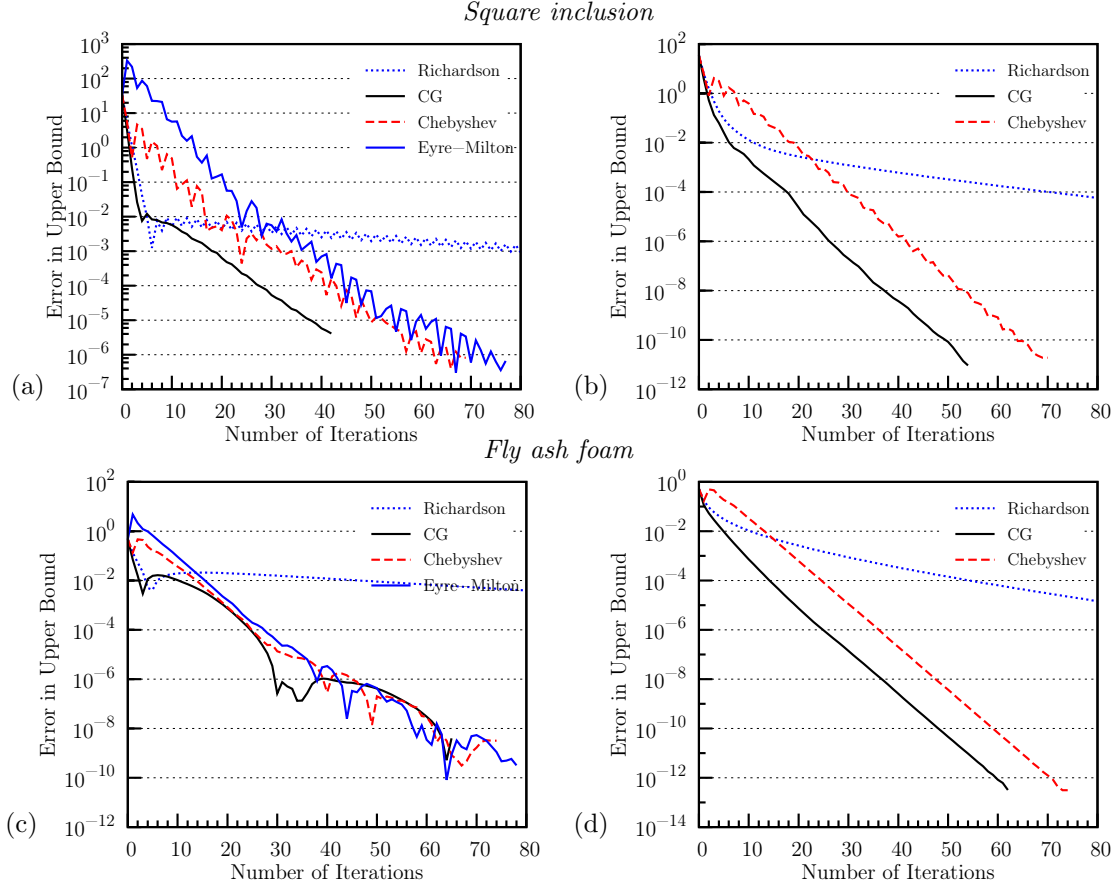
Figure 6: Convergence of error in guaranteed upper bound for GaNi (left) and Ga (right) discretizations.

4.6. Condition number

To conclude our numerical experiments, in Figure 11 we plot the dependence of the number of iterations to reach the accuracy of $\epsilon = 10^{-6}$ on the condition number of system matrix κ. Note that we present the results only for the square inclusion, Figure 1(a), and set the coefficient of the inclusion equal to κ (analogous conclusions hold for more complex microstructures). In addition, all algorithms are terminated by the relative residual norm.

The results generally follow the trend predicted by the error estimates (21), (23), and (25) and confirm the linear scaling $\mathcal{O}(\kappa)$ for the Richardson scheme and $\mathcal{O}(\sqrt{\kappa})$ for the remaining solvers. Closer inspection reveals that some solvers converge *faster* for large values of κ, namely the Conjugate gradient method or the Richardson scheme for Ga discretization. Such superconvergent behavior was reported in earlier studies by, e.g., Monchiet and Bonnet [16] and Willot et al. [18] for FFT-based solvers, and Schneider et al. [57] for finite difference methods, in order to demonstrate that their methods work for $\kappa \to \infty$. On the basis of the presented results, one may thus conjecture that iterative methods applied to Ga-based systems will converge for the infinite contrast of material coefficients. We would like to address this question in our future work, together with a more detailed investigation into the distribution of discretization errors.

5. Conclusions

In this paper, we have performed a comparative study of iterative algorithms for systems of linear equations arising from a Fourier-Galerkin discretization of the periodic corrector problem. Two discretization schemes have been considered, exact (Ga) and trapezoidal (GaNi) integration, and the ensuing systems of linear equations have been solved with the Richardson, the Chebyshev, the Conjugate gradient, and the Eyre-Milton algorithms.

16
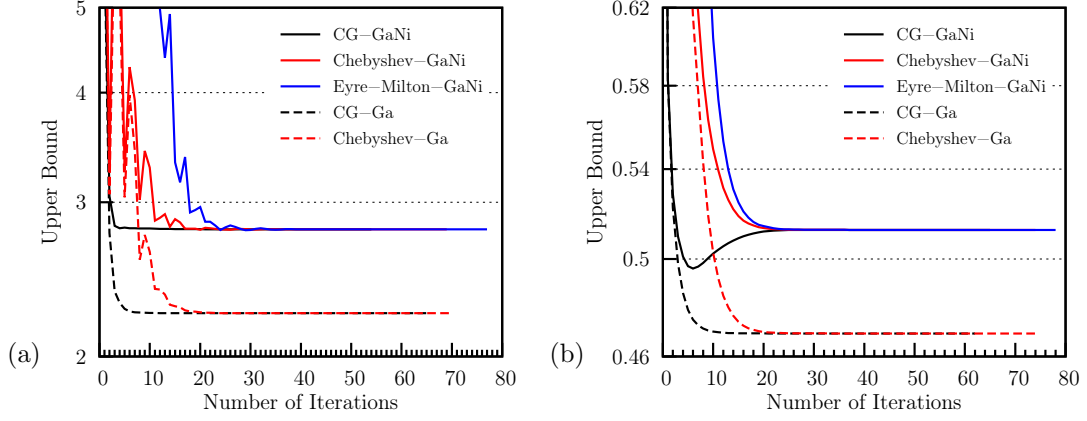
Figure 7: Effect of numerical integration on tightness of guaranteed upper bound for (a) square inclusion and (b) fly ash foam microstructures.
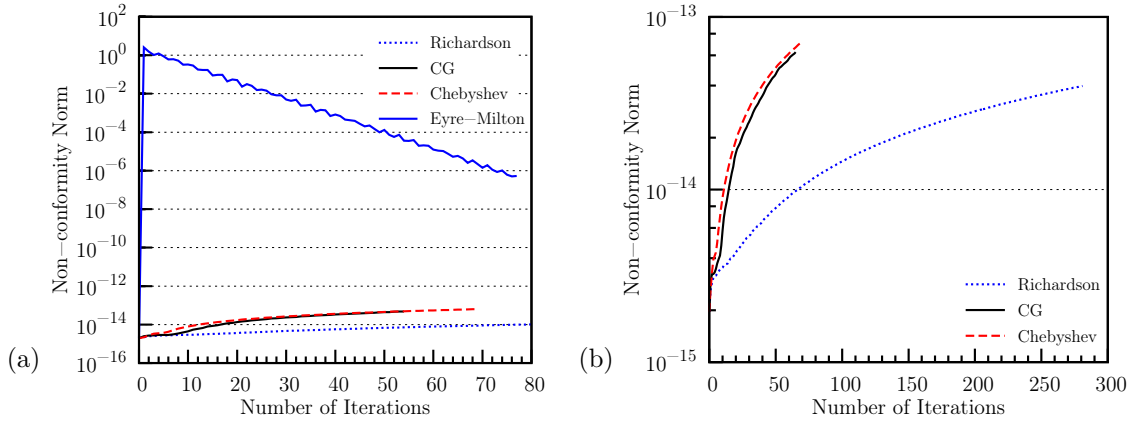
Figure 8: Evolution of non-conformity norm during iterations for square inclusion and (a) GaNi and (b) Ga discretizations.

Based on the outcomes of our study, summarized in Table 1 for the reader's convenience, we conclude that:

(i) In terms of the *rate of convergence*, the Conjugate gradient, the Chebyshev, and the Eyre-Milton algorithms exhibit super-linear convergence and the Richardson method converges with the linear rate. In addition, the Conjugate gradient method appears to be the most efficient solver, while the Chebyshev and the Eyre-Milton algorithms display comparable performance.

(ii) All three general-purpose solvers — the Richardson, the Conjugate gradient, and the Chebyshev algorithms — generate iterates that *conform* to $\mathbb{E}$, the space associated with curl-free and zero-mean trigonometric polynomials. The Eyre-Milton method produces non-conforming iterates.

(iii) The general-purpose solvers work for linear systems arising from both Ga and GaNi discretizations, while the Eyre-Milton method is applicable exclusively to the GaNi setting.

(iv) The approximate upper bounds generated by the Richardson and the Conjugate gradient methods for the Ga discretizations exhibit *monotone convergence*, whereas all other options yield non-monotone convergence. The superior performance of the Conjugate gradient algorithm follows from the fact that the bound corresponds to the energy norm that Conjugate gradients naturally minimize.

(v) With regard to *memory efficiency* of the implementations introduced in Section 3, the Conjugate gradients need to store one additional vector per iteration. The computational

17
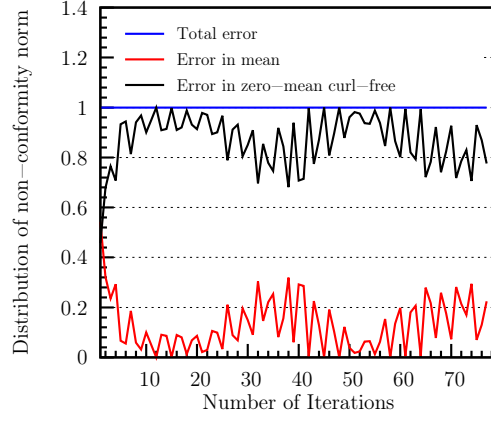
Figure 9: Relative distribution of non-conformity norm in the Eyre-Milton algorithm for the square inclusion problem and GaNi discretization.
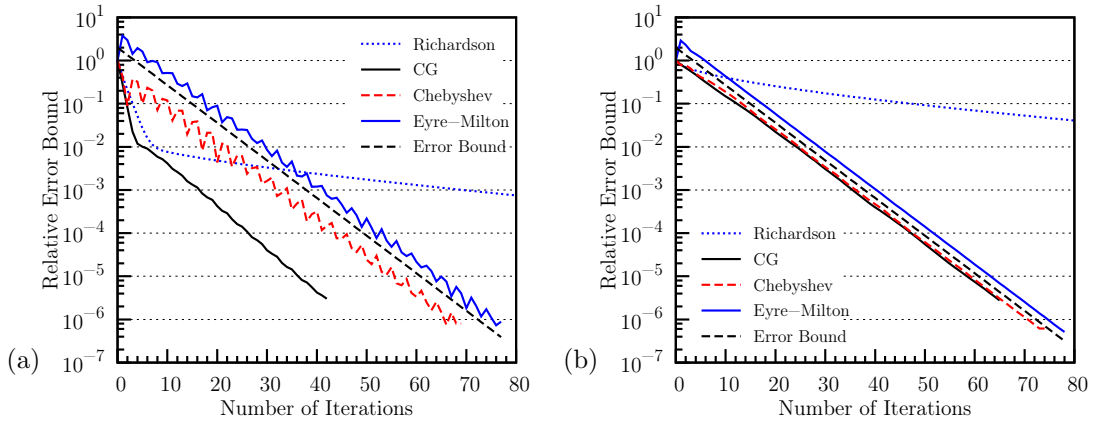
Figure 10: Performance of error estimates for GaNi discretization and (a) square inclusion and (b) fly ash foam microstructures; error bound is defined by (23).

complexity of a single iteration of all algorithms is comparable because it is dominated by the forward and inverse FFTs.

We believe that the computational observations collected in this paper provide a convenient starting point for the development of even more efficient solvers for FFT-based homogenization algorithms. Our particular interest is to continue to explore the potential of the Chebyshev method to achieve a more robust convergence, e.g. [58], or to serve as a preconditioner in multi-grid solvers [59]. Investigations of these topics are underway and results will be reported separately.

Acknowledgments

Figure 11: Dependence of the number of iterations to convergence on the condition number of the system matrix for square inclusion and (a) GaNi and (b) Ga discretizations.

Algorithm	Scaling	Conforming	Ga	Upper bound	Storage
Richardson	$\mathcal{O}(\kappa^1)$	Yes	Yes	Monotone	3
Conjugate gradients	$\mathcal{O}(\kappa^{\frac{1}{2}})$	Yes	Yes	Monotone	4
Chebyshev	$\mathcal{O}(\kappa^{\frac{1}{2}})$	Yes	Yes	Non-monotone	3
Eyre-Milton	$\mathcal{O}(\kappa^{\frac{1}{2}})$	No	No	Non-monotone	3

Table 1: Overall comparison of the four short-recurrence iterative algorithms; *scaling* indicates how the number of iterations to converge grows with an increasing condition number of system matrix κ, *conforming* algorithms generate iterates from subspace $\mathbb{E}$, Ga refers to the extendibility of the algorithm to the exact integration, *upper bound* indicates how the approximate upper bounds generated by the iterates converge to the guaranteed upper bound, *storage* specifies the number of vectors needed in one iteration.

References

[1] M. D. Uchic, M. A. Groeber, A. D. Rollett, Automated serial sectioning methods for rapid collection of 3-D microstructure data, JOM 63 (3) (2011) 25–29. doi:10.1007/s11837-011-0041-2.

[2] E. Maire, P. J. Withers, Quantitative X-ray tomography, International Materials Reviews 59 (1) (2013) 1–43. doi:10.1179/1743280413Y.0000000023.

[3] F. Ballani, D. Stoyan, Reconstruction of random heterogeneous media, Journal of Microscopy 258 (3) (2015) 173–178. doi:10.1111/jmi.12234.

[4] B. Sonon, B. François, T. J. Massart, An advanced approach for the generation of complex cellular material representative volume elements using distance fields and level sets, Computational Mechanics 56 (2) (2015) 221–242. doi:10.1007/s00466-015-1168-8.

[5] D. Cioranescu, P. Donato, An Introduction to Homogenization, Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 1999.

[6] G. W. Milton, The Theory of Composites, Vol. 6 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2002. doi:10.1017/CBO9780511613357.

[7] M. G. D. Geers, V. G. Kouznetsova, W. A. M. Brekelmans, Multi-scale computational homogenization: Trends and challenges, Journal of Computational and Applied Mathematics 234 (7) (2010) 2175–2182. doi:10.1016/j.cam.2009.08.077.

[8] K. Terada, T. Miura, N. Kikuchi, Digital image-based modeling applied to the homogenization analysis of composite materials, Computational Mechanics 20 (4) (1997) 331–346. `doi:10.1007/s004660050255`.

[9] E. Garboczi, Finite element and finite difference programs for computing the linear electric and elastic properties of digital images of random materials., Tech. Rep. NISTIR 6269, Building and Fire Research Laboratory, National Institute of Standards and Technology, Gaithesburg, Maryland 2089 (1998).

[10] D. H. Pahr, P. K. Zysset, From high-resolution CT data to finite element models: development of an integrated modular framework, Computer Methods in Biomechanics and Biomedical Engineering 12 (1) (2009) 45–57. `doi:10.1080/10255840802144105`.

[11] H. Moulinec, P. Suquet, A fast numerical method for computing the linear and nonlinear mechanical properties of composites, Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie 318 (11) (1994) 1417–1423.

[12] D. J. Eyre, G. W. Milton, A fast numerical scheme for computing the response of composites using grid refinement, The European Physical Journal Applied Physics 6 (1) (1999) 41–47. `doi:10.1051/epjap:1999150`.

[13] J.-C. Michel, H. Moulinec, P. Suquet, A computational method based on augmented Lagrangians and fast Fourier transforms for composites with high contrast, CMES: Computer Modeling in Engineering & Sciences 1 (2) (2000) 79–88. `doi:10.3970/cmes.2000.001.239`.

[14] J.-C. Michel, H. Moulinec, P. Suquet, A computational scheme for linear and non-linear composites with arbitrary phase contrast, International Journal for Numerical Methods in Engineering 52 (1–2) (2001) 139–160. `doi:10.1002/nme.275`.

[15] V. Vinogradov, G. W. Milton, An accelerated FFT algorithm for thermoelastic and nonlinear composites, International Journal for Numerical Methods in Engineering 76 (11) (2008) 1678–1695. `doi:10.1002/nme`.

[16] V. Monchiet, G. Bonnet, A polarization-based FFT iterative scheme for computing the effective properties of elastic composites with arbitrary contrast, International Journal for Numerical Methods in Engineering 89 (11) (2012) 1419–1436. `doi:10.1002/nme.3295`.

[17] G. Bonnet, Effective properties of elastic periodic composite media with fibers, Journal of the Mechanics and Physics of Solids 55 (5) (2007) 881–899. `doi:10.1016/j.jmps.2006.11.007`.

[18] F. Willot, B. Abdallah, Y.-P. Pellegrini, Fourier-based schemes with modified Green operator for computing the electrical response of heterogeneous media with accurate local fields, International Journal for Numerical Methods in Engineering 98 (7) (2014) 518–533. `doi:10.1002/nme.4641`.

[19] F. Willot, Fourier-based schemes for computing the mechanical response of composites with accurate local fields, Comptes Rendus Mécanique 343 (2015) 232–245. `doi:10.1016/j.crme.2014.12.005`.

[20] L. Gélébart, F. Ouaki, Filtering material properties to improve FFT-based methods for numerical homogenization, Journal of Computational Physics 294 (2015) 90–95. `doi:10.1016/j.jcp.2015.03.048`.

[21] M. Kabel, D. Merkert, M. Schneider, Use of composite voxels in FFT-based homogenization, Computer Methods in Applied Mechanics and Engineering 294 (2015) 168–188. `doi:10.1016/j.cma.2015.06.003`.

[22] J. Zeman, J. Vondřejc, J. Novák, I. Marek, Accelerating a FFT-based solver for numerical homogenization of periodic media by conjugate gradients, Journal of Computational Physics 229 (21) (2010) 8065–8071. `doi:10.1016/j.jcp.2010.07.010`.

[23] J. Vondřejc, J. Zeman, I. Marek, An FFT-based Galerkin method for homogenization of periodic media, Computers & Mathematics with Applications 68 (3) (2014) 156–173. `doi:10.1016/j.camwa.2014.05.014`.

[24] M. Schneider, Convergence of FFT-based homogenization for strongly heterogeneous media, Mathematical Methods in the Applied Sciences 38 (13) (2015) 2761–2778. `doi:10.1002/mma.3259`.

[25] S. Brisard, L. Dormieux, Combining Galerkin approximation techniques with the principle of Hashin and Shtrikman to derive a new FFT-based numerical method for the homogenization of composites, Computer Methods in Applied Mechanics and Engineering 217–220 (2012) 197–212. `doi:10.1016/j.cma.2012.01.003`.

[26] S. Brisard, F. Legoll, Periodic homogenization using the Lippmann–Schwinger formalism`arXiv:1411.0330`.

[27] J. Vondřejc, Improved guaranteed computable bounds on homogenized properties of periodic media by the Fourier–Galerkin method with exact integration, International Journal for Numerical Methods in Engineering`doi:10.1002/nme.5199`.

[28] J. Vondřejc, J. Zeman, I. Marek, Guaranteed upper–lower bounds on homogenized properties by FFT-based Galerkin method, Computer Methods in Applied Mechanics and Engineering 297 (2015) 258–291. `doi:10.1016/j.cma.2015.09.003`.

[29] V. Monchiet, Combining FFT methods and standard variational principles to compute bounds and estimates for the properties of elastic composites, Computer Methods in Applied Mechanics and Engineering 283 (2015) 454–473. `doi:10.1016/j.cma.2014.10.005`.

[30] L. Richardson, The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam, Philosophical Transactions of the Royal Society A 210 (1911) 307–357. `doi:10.1098/rsta.1911.0009`.

[31] M. R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Bureau of Standards 49 (6) (1952) 409–463.

[32] C. Lanczos, Chebyshev polynomials in the solution of large-scale linear systems, in: Proceedings of the Association for Computing Machinery, Toronto, 1952(1953), Sauls Lithograph Co., Washington D. C., 1953, pp. 124–133.

[33] H. Moulinec, F. Silva, Comparison of three accelerated FFT-based schemes for omputing the mechanical response of composite materials, International Journal for Numerical Methods in Engineering 97 (13) (2014) 960–985. `doi:10.1002/nme.4614`.

[34] J.-C. Michel, H. Moulinec, P. Suquet, Effective properties of composite materials with periodic microstructure: a computational approach, Computer Methods in Applied Mechanics and Engineering 172 (1–4) (1999) 109–143. `doi:10.1016/S0045-7825(98)00227-8`.

[35] A. Prakash, R. A. Lebensohn, Simulation of micromechanical behavior of polycrystals: finite elements versus fast Fourier transforms, Modelling and Simulation in Materials Science and Engineering 17 (6) (2009) 064010. `doi:10.1088/0965-0393/17/6/064010`.

[36] B. Liu, D. Raabe, F. Roters, P. Eisenlohr, R. A. Lebensohn, Comparison of finite element and fast Fourier transform crystal plasticity solvers for texture prediction, Modelling and Simulation in Materials Science and Engineering 18 (8) (2010) 85005. `doi:10.1088/0965-0393/18/8/085005`.

[37] C. Robert, C. Mareau, A comparison between different numerical methods for the modeling of polycrystalline materials with an elastic–viscoplastic behavior, Computational Materials Science 103 (2015) 134–144. `doi:10.1016/j.commatsci.2015.03.028`.

[38] C. F. Dunant, B. Bary, A. B. Giorla, C. Péniguel, J. Sanahuja, C. Toulemonde, A.-B. Tran, F. Willot, J. Yvonnet, A critical comparison of several numerical methods for computing effective properties of highly heterogeneous materials, Advances in Engineering Software 58 (2013) 1–12. `doi:10.1016/j.advengsoft.2012.12.002`.

[39] B. Bary, L. Gélébart, E. Adam, C. Bourcier, Numerical analysis of linear viscoelastic 3D concrete specimens: Comparison between FE and FFT methods, in: N. Bicanić, H. Mang, G. Meschke, R. de Borst (Eds.), Computational Modelling of Concrete Structures - Proceedings of EURO-C 2014, CRC Press/Balkema, 2014, pp. 373–381.

[40] H. Moulinec, P. Suquet, Comparison of FFT-based methods for computing the response of composites with highly contrasted mechanical properties, Physica B: Condensed Matter 338 (1–4) (2003) 58–60. `doi:10.1016/S0921-4526(03)00459-9`.

[41] J. Vondřejc, J. Zeman, I. Marek, Analysis of a Fast Fourier transform based method for modeling of heterogeneous materials, in: I. Lirkov, S. Margenov, J. Wasniewski (Eds.), Large-Scale Scientific Computing, Vol. 7116 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2012, pp. 515–522. `doi:10.1007/978-3-642-29843-1_58`.

[42] P. Shanthraj, P. Eisenlohr, M. Diehl, F. Roters, Numerically robust spectral methods for crystal plasticity simulations of heterogeneous materials, International Journal of Plasticity 66 (2015) 31–45. `doi:10.1016/j.ijplas.2014.02.006`.

[43] T. W. J. de Geus, J. Vondřejc, J. Zeman, R. H. J. Peerlings, M. G. D. Geers, Finite strain FFT-based non-linear solvers made simple`arXiv:1603.08893`.

[44] J. Zeman, T. W. J. de Geus, J. Vondřejc, R. H. J. Peerlings, M. G. D. Geers, A finite element perspective on non-linear FFT-based micromechanical simulations`arXiv:1601.05970`.

[45] J. Saranen, G. Vainikko, Periodic Integral and Pseudodifferential Equations with Numerical Approximation, Springer Monographs Mathematics, Berlin, Heidelberg, 2002. `doi:10.1007/978-3-662-04796-5`.

[46] R. Luciano, E. Sacco, Variational methods for the homogenization of periodic heterogeneous media, European Journal of Mechanics - A/Solids 17 (4) (1998) 599–617. `doi:10.1016/S0997-7538(99)80024-2`.

[47] D. Young, On Richardson's method for solving linear systems with positive definite matrices, Journal of Mathematical Physics 32 (1954) 243–255. `doi:10.1002/sapm1953321243`.

[48] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2003. `doi:10.1137/1.9780898718003`.

[49] J. W. Daniel, The conjugate gradient method for linear and nonlinear operator equations, SIAM Journal on Numerical Analysis 4 (1967) 10–26. `doi:10.1137/0704002`.

[50] P. Concus, G. H. Golub, D. P. O'Leary, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, in: Sparse matrix computations (Proc. Sympos., Argonne Nat. Lab., Lemont, Ill., 1975), Academic Press, New York, 1976, pp. 309–332.

[51] S. Brisard, L. Dormieux, FFT-based methods for the mechanics of composites: A general variational framework, Computational Materials Science 49 (3) (2010) 663–671. doi:10.1016/j.commatsci.2010.06.009.

[52] G. H. Golub, R. S. Varga, Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. I, Numerische Mathematik 3 (1961) 147–156. doi:10.1007/BF01386013.

[53] M. H. Gutknecht, S. Röllin, The Chebyshev iteration revisited, Parallel Computing 28 (2) (2002) 263–283. doi:10.1016/S0167-8191(01)00139-9.

[54] T. Gergelits, Z. Strakoš, Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations, Numerical Algorithms 65 (4) (2014) 759–782. doi:10.1007/s11075-013-9713-z.

[55] P. Hlaváček, V. Šmilauer, F. Škvára, L. Kopecký, R. Šulc, Inorganic foams made from alkali-activated fly ash: Mechanical, chemical and physical properties, Journal of the European Ceramic Society 35 (2) (2015) 703–709. doi:10.1016/j.jeurceramsoc.2014.08.024.

[56] J. Papež, J. Liesen, Z. Strakoš, Distribution of the discretization and algebraic error in numerical solution of partial differential equations, Linear Algebra and its Applications 449 (2014) 89–114. doi:10.1016/j.laa.2014.02.009.

[57] M. Schneider, F. Ospald, M. Kabel, Computational homogenization of elasticity on a staggered grid, International Journal for Numerical Methods in Engineering 105 (9) (2016) 693–720. doi:10.1002/nme.5008.

[58] A. Wathen, T. Rees, Chebyshev semi-iteration in preconditioning for problems including the mass matrix, Electronic Transactions on Numerical Analysis 34 (2008/09) 125–135.

[59] M. Adams, M. Brezina, J. Hu, R. Tuminaro, Parallel multigrid smoothing: polynomial versus Gauss-Seidel, Journal of Computational Physics 188 (2) (2003) 593–610. doi:10.1016/S0021-9991(03)00194-3.